



Assessing Secondary School Students' Justifications for Supporting or Rejecting a Scientific Hypothesis in the Physics Lab

Tobias Ludwig¹  · Burkhard Priemer² · Doris Lewalter³

Published online: 1 June 2019

© The Author(s) 2019

Abstract

Justifications play a central role in argumentation, which is a core topic in school science education. This paper contributes to this field of research by presenting two studies in which we assess students' justifications for supporting or rejecting hypotheses in the physics lab based on self-collected, anomalous experimental data, which are defined as data that contradict a prior belief, hypothesis, or concept. Study 1 analyzes the spectrum of possible justifications students give in semi-structured interviews and categorizes these into ten types: *appeal to an authority*, *data as evidence*, *experimental competence (technical/skills)*, *experimental competence (self-concept)*, *ignorance*, *intuition*, *measurement uncertainties (explicit)*, *measurement uncertainties (implicit)*, *suitability of the experimental setup*, and *use of theoretical concepts*. Study 2 presents a questionnaire suitable for medium- and large-scale assessments that probes students' use of four of these types of justifications: *appeal to an authority*, *data as evidence*, *intuition*, and *measurement uncertainties (explicit)*. The questionnaire can be administered in 5–10 minutes and is designed for students in the eighth and ninth grades. We outline the development and quality of the assessment tools of both studies, reporting on the content validity, factorial validity, discriminant validity, convergent validity, and reliability of the questionnaire. The two studies shed light on the various justifications students use when evaluating anomalous data at a fine-grained level.

Keywords Argumentation · Justification · Data · Experimentation · Hypotheses · Physics

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11165-019-09862-4>) contains supplementary material, which is available to authorized users.

✉ Tobias Ludwig
tobias.ludwig@ph-karlsruhe.de

Burkhard Priemer
priemer@physik.hu-berlin.de

Doris Lewalter
doris.lewalter@tum.de

Extended author information available on the last page of the article

In science education, analyzing and producing scientific arguments are important processes in “doing science” because they draw on conceptual understanding rather than on rote memorization of facts (Driver et al. 2000). In the production of an argument, one key factor is the justification of why given information or data support a certain claim. Thus, competences in argumentation, and in justification in particular, are implemented in different science standards (US: NGSS Lead States 2013, Appendix F; UK: Department for Education 2015, e.g., Upper key stage 2; Germany: KMK 2004, p. 9). However, learning to argue is a complex and demanding task that is not easily achieved (Katchevich et al. 2013; Osborne et al. 2013; Sandoval and Millwood 2005). Sampson and Clark (2008) reported that “when students do include a justification in an argument, they tend to rely on inferences, personal experiences, and authority figures” (p. 468). Thus, it is necessary to identify the types of justifications that learners use when they are engaged in scientific argumentation. However, assessing students’ use of arguments is a demanding task and a future challenge for science education research (Henderson et al. 2018). This paper contributes to this research by analyzing students’ situation-specific justifications in the context of laboratory work. In contrast to other work done in this field, we focus on high school students conducting a quantitative physics experiment that produces anomalous measurement data, which are defined here as data that contradict a prior belief, hypothesis, or concept. Here, we elicit, examine, and categorize the justifications students use when choosing to support or reject a scientific hypothesis regardless of whether they undergo conceptual change with respect to the physics topic. Furthermore, we present a questionnaire that is suitable not only for small sample sizes but also for medium- and large-scale assessments that probe, in a valid and reliable manner, students’ use of justifications in a lab setting. Such a paper-and-pencil tool for assessing justifications in a short time (5–10 minutes) has not previously been developed. This work contributes to the existing body of research describing the nature of justification, that is, “how ideas or claims are supported or validated within an argument” (Sampson and Clark 2008, p. 449), by analyzing students’ use of justification in the context of learning physics. Furthermore, an assessment tool for justifications suitable for use with larger sample sizes can help in investigations of the personal and situational factors that influence the use of certain types of justifications and of which types of justifications support students’ understanding of scientific concepts.

Theoretical Framework

When investigating justifications in arguments based on empirical experimental data to support or reject a hypothesis in physics, we must review the existing research on students’ use of justifications in general. This allows us to locate our work in the context of learning situations where certain claims are given, and typical data sets are collected by students. Because our focus is on quantitative experiments, we further review the research on argumentation in lab work settings. That informs us about how to trigger and analyze argumentation in these learning environments. Finally, we outline research on students’ reasoning when supporting or rejecting a scientific hypothesis in the light of anomalous data, which we used to initiate argumentation. These references are central to showing how our work differs from other work conducted in the field; for example, we examine students’ justifications regardless of whether they change their belief or concept about a certain topic in physics. Before describing this theoretical background, however, we will briefly define some basic terms to prevent misunderstandings.

We use the term “argument” to denote “artifacts that a student or a group of students create when asked to articulate and justify claims or explanations” (Sampson and Clark 2008, p. 447).

These artifacts are developed from elements such as claims, data, and warrants, as described by Toulmin (2003). By referring to (a combination of) these elements, an argument can be described more precisely: “We mean the putting forward of reasons where claims are justified by relating them to the data on which they are based” (Osborne et al. 2001, p. 63). In the process of scientific inquiry, a claim can be a hypothesis about a relation between physical quantities. We use the term “hypothesis” to denote students’ assumptions about a relation between quantities in physics, for example, the influence of the mass of a bob on the oscillation time of a simple pendulum (Kanari and Millar 2004). More precisely, we focus on hypotheses regarding (physics) concepts and follow the interpretation of Scott et al. (2007) who characterize a concept as a “knowledge unit” (p. 36). Using this interpretation, a claim or hypothesis for a given context in physics can be considered a concept. In our studies, data are considered both observations of scientific experiments (e.g., observing a mass swinging on a string and counting time intervals) and quantitative measurements (e.g., sets of numbers for oscillation times).

The Nature of Justification in Argument

There are different analytical frameworks to investigate argument and argumentation in science education. Sampson and Clark (2008, p. 447) differentiate these approaches by means of three different “focal issues: structure, justification and content.” While analyzing the structure and content of arguments is obviously an important matter in science education, analyzing the nature of justifications is particularly enlightening in the context of lab work because it allows research to gain insight into *how* students generate knowledge from scientific inquiry. We want to further advance this argument in the following paragraph.

In science, and especially in science education, knowledge construction is linked to the justification of knowledge by relating scientific claims to data (Jiménez-Aleixandre and Erduran 2007, p. 13). Consequently, justifications can be seen as information components of an argument that shows *why* data counts as evidence to support a claim (McNeill and Krajcik 2007; Ryu and Sandoval 2012). Analyzing the nature of justifications allows us to focus on the “epistemic dimension” in argument (Weinberger and Fischer 2006, p. 74) and, consequently, learn about *how* students construct knowledge in argumentation and *why* they support or reject a certain claim. This stance contrasts with research based on Toulmin’s work, which instead analyzes the structural components of an argument. The nature of justification in argument has previously been addressed in research (Kelly et al. 1998; Sandoval 2003; Sandoval and Millwood 2005; Zohar and Nemet 2002). Chinn et al. (2011, p. 154) summarized that the focus of research is on epistemic beliefs about types of justifications, such as referring to authority, experience, personal views, rationality, and rules of inquiry. These types of justifications in prior research informed our work about the reasons students give for supporting a claim. We draw on these results in our data analysis. However, our work is accomplished in the context of lab work, where experiments can influence students’ justifications. Hence, we expect justification types that are specific to this learning environment. Before we outline research about argumentation in lab work settings, we note some constraints in investigating students’ justifications, which consequently influenced our work.

First, we restrict our view on argument and especially on justifications to educational purposes because a good argument or justification in science is not necessarily a good argument or justification in the general school context (Sandoval 2003): Decisions regarding whether there is enough evidence for a claim or what counts as good evidence can be different in school inquiry tasks than in scientific research. Our work focuses on school learning and hence does not make inferences about scientists’ justifications.

Further, justifications depend on the field (Toulmin 2003) and on the cultural context (Ryu and Sandoval 2012) or situation (Chinn et al. 2011; Sandoval 2012) in which they are made. This means that the justifications students give in a physics class may differ from justifications they would give in a chemistry class, even if the nature of the data (e.g., anomalous) is the same and the experiments in both disciplines are comparable. Moreover, a student may give different justifications for the same topic in different situations (Chinn et al. 2011), for example, when the personal relevance of the topic has changed. This limits the generalizability of many studies, including our own, because we used only two physics contexts. However, the debate about the domain-specificity of argumentation and reasoning is ongoing: Kind (2013, p. 534) indicates that while reasoning is knowledge-dependent—and thus related to domains—domain-general strategies also exist.

Finally, justifications are subject to biases, such as confirmation bias or the fear of rejection (Ganser and Hammann 2009; Klayman and Ha 1987; Njoo and de Jong 1993; van Joolingen and de Jong 1997). An inadequate analysis of experimental data, in this case, is not caused by a lack of logical thinking, but rather, by attempts to confirm prior existing concepts (Hammann et al. 2006; Schulz et al. 2008). In our studies, the categorizations of students' justifications were developed so that these prior concepts can be identified and included.

The Use of Justifications in Laboratory Work

We frame our work in the context of a school laboratory. A number of studies relevant to lab work have focused on argumentation based on experimental data (Jeong et al. 2007; Kanari and Millar 2004; Kind et al. 2011; Lubben et al. 2010) and dealing with hypotheses (Klayman and Ha 1987; Njoo and de Jong 1993; van Joolingen and de Jong 1997). Discussion is ongoing regarding whether lab work in general is suitable to trigger argumentation (Katchevich et al. 2013, p. 338). However, research indicates that while the quality of arguments derived from typical experiments is low (Kind et al. 2011, p. 2530), inquiry-type experiments (guided or open) seem to have the potential to serve as platforms on which to construct arguments (Duschl and Osborne 2002; Katchevich et al. 2013; Wilson et al. 2010). Two important conditions of argument-rich learning environments were found: (a) students should be at the center of the process (Jiménez-Aleixandre 2007) and (b) students should familiarize themselves with the experiment and generate their own data (Kind et al. 2011, p. 2528). We draw from these results by developing a lab work task that poses an open question (Does the mass of the bob of a simple pendulum have an influence on the time of oscillation?) and letting students investigate their own hypothesis by running experiments by themselves and collecting their own measurement data. Our pendulum example is created to generate data that is “astonishing” (e.g., not expected) for most of the students. This can cause a debate or cognitive conflict (Kind et al. 2011, p. 2531). Kind and colleagues looked at how different forms of data and tasks stimulate debate. Their focus was on identifying elements of argumentation (Toulmin 2003) in debates about the results of experiments (e.g., bodies changing temperature, dissolving salt in water). In contrast, our studies focus on justifications rather than the structure of arguments and on students working by themselves rather than in groups. It is true that argumentation is often socially embedded (Kolstø and Ratcliffe 2007). We, as scientists, argue *with* someone with the aim of persuasion (Ryu and Sandoval 2012). However, argumentation is not necessarily dialogical; it can also be considered from an individual perspective (Driver et al. 2000; Jiménez-Aleixandre and Erduran 2007). This is, for example, the case when arguments are presented in written texts or oral statements, as in our work. Here, the statements reflect intramental sense making, which is part of the individual processes that contribute to constructing

arguments (Ford 2012). How this intra-mental form of debate influences students' justifications in lab work settings is as yet unknown. On one hand, students who work by themselves cannot make use of their peers' ideas and arguments. This may limit their justifications. On the other hand, these students cannot be dominated by peers who guide the lab work investigation and may overlook others' justifications. This is especially important when the results of an experiment contradict students' expectations. In contrast to the work cited, which analyzed group discussions, we looked at students' justifications when they work independently.

Justifications When Supporting or Rejecting a Hypothesis

Justifying claims or hypotheses about scientific concepts is related to conceptual understanding (McDonald and Kelly 2012) and—when scientific concepts do not match students' every-day life experiences—conceptual change (Asterhan and Schwarz 2009). Students hold views about scientific concepts that may contradict scientific interpretations. These preconceptions are field- and context-dependent and may even be logically incoherent (Hammer and Elby 2002). The literature on conceptual change further indicates that students' understanding—even though it may not be scientifically correct—can be a valuable source for acquiring extended scientific understanding (Brown and Hammer 2008). Posner et al. (1982) provided a well-known list of factors that influence conceptual change. One particularly important finding is that a concept may be changed based on non-rational arguments, which is described by Sinatra (2005) as “hot conceptual change.” For example, a person's affect (Gregoire 2003) and motivation and the social context of a message (Dole and Sinatra 1998) determine whether an existing concept will be discarded. This is important for our work, as we analyze students' justifications that may also be non-rationally driven and may address empirical results that are not in line with their expectations. The need to build an argument could be triggered by controversial information such as surprising effects, observations that contradict expectations, and data that are seen as “astonishing.” Thus, we follow Chinn and Brewer (1998) and others (Chan et al. 1997; Katchevich et al. 2013, p. 319; Kind et al. 2011; Shepardson 1999) by using anomalous data as a source to provoke cognitive dissonance and generate argumentation in order to fulfill the condition of “dissatisfaction with the current concept” described by Posner et al. (1982). In a broader sense, anomalous data can be any kind of information; for example, texts explaining studies and theories that challenge or contradict students' views on certain topics, such as the extinction of dinosaurs (Chinn and Brewer 1998) or the construction of the great pyramids in Giza, Egypt (Mason 2001). We focus here on more specific empirical observation and measurement data from students' *own* laboratory experiments. Research shows that there are multiple ways in which students respond to or—within our context—argue with anomalous data: Kanari and Millar (2004) demonstrate that learners do not initially consider unclear data and will adhere to familiar concepts that are based on previous knowledge or guesses. Only after a certain degree of unambiguity are learners ready to contemplate data and potentially change their conception. Accordingly, Chinn and Malhotra (2002) report that “[...] less ambiguous data promoted some change in beliefs, whereas more ambiguous anomalous data did not” (p. 333). Learners show great skill in rejecting conflicting evidence in order to retain their own hypotheses (Robinson 2000).

However, which justifications do students give in light of “astonishing” data? Abi-El-Mona and Abd-El-Khalick (2011) investigated students' justifications in debates and interviews about qualitative experiments in chemistry. They observed that “warrants were loosely based on prior experience or course content knowledge, but did explicitly and clearly link claims to the target observations and/or justifications” (p. 354). In their analysis, Abi-El-Mona and Abd-El-Khalick (2011, p. 355) refer to Walton's (1996) reasoning scheme, which discriminates among argument

from sign, argument from example, argument from verbal classification, argument from expert opinion, argument from evidence to hypothesis, and argument from analogy. With the possible exception of referring to experts, all justifications are rational and focus on either theory or observation. Kelly et al. (2001, p. 167) found that in addition to trusting an author, students ignore information. This research indicated to us that we should capture both rational and non-rational justifications. Chinn and Brewer (1998) identified eight possible responses to anomalous data in science containing justifications: (a) ignoring the data, (b) rejecting the data, (c) professing uncertainty about the validity of the data, (d) excluding the data from the domain of the current theory, (e) holding the data in abeyance, (f) reinterpreting the data, (g) accepting the data and making peripheral changes to the current theory, and (h) accepting the data and changing theories.

Lin (2007) allowed participants to respond to anomalous data in more than one category at the same time and added the category “uncertainty about interpretation” to the Chinn and Brewer’s (1998) categories. We support this view, as students may have multiple justifications when accepting or declining a claim or hypothesis. Further, we challenge the general validity of the Chinn and Brewer’s (1998) categories for science education for the following reasons. (a) The use of one’s *own* experimental data is missing; Chinn and Brewer (1998) provide participants with theories followed by anomalous data. The source and trustworthiness of the data remain unclear. Hence, responses to anomalous data may be different if students collect the data themselves. There is empirical evidence for this claim, although previous studies do not lead to unambiguous results: While Delen and Krajcik (2015) showed that students using first-hand data might generate stronger explanations, Kanari and Millar (2004) report that there is no difference in reasoning between students who recorded video experiments and students who carried out the experiment themselves. In addition, knowledge construction is facilitated by one’s own experimental work versus being provided with information (Anderson 2007). Thus, we can conclude that the resulting cognitive activities will be different between settings with first- or second-hand data (Hug and McNeill 2008). (b) Relations to every-day life experiences are missing; in Chinn and Brewer’s (1998), Mason’s (2001), and Lin’s (2007) studies, the contexts do not overlap with participants’ own experiences. However, responses to anomalous data in terms of accepting or declining a hypothesis may be influenced by students’ preconceptions (Posner et al. 1982). Espinoza and Quarless (2010) report that students can arrive at correct and incorrect solutions after having made incorrect predictions, in which case the learning outcome was influenced only by students’ abilities to judge the uncertainties in the data at hand. (c) There is no classification for students’ responses to anomalous data categorized independently from the students’ decision to maintain or change their hypothesis; Chinn and Brewer (1998) categorize these responses in light of maintaining or changing an initial concept. However, students may ignore or reject or reinterpret data in both cases—keeping or changing their initial hypothesis or concept. These three reasons suggest that responses to anomalous data are influenced by factors that were not examined by Chinn and Brewer (1998). Having said that, we do not argue that the categorizations of Chinn and Brewer (1998) and Mason (2001) do not hold; we simply argue that they may have limited applicability regarding learning in school labs.

Research Questions

In this paper, we want to bridge the research on argumentation in lab work settings and the research on students’ responses to self-generated anomalous data. As outlined above, this intersection is currently only rarely covered by previous research: So far, very little is known on how students justify a hypothesis based on their own anomalous data from science experiments that consider these

three aspects listed above. We want to contribute to filling this research gap in two ways: First, there is an overall need to categorize students' justifications in light of self-generated anomalous data. This will allow researchers and practitioners to better understand students' epistemic processes in the lab. Here, we aim to capture a wide spectrum of possible justifications in a first step. Second, we follow Chinn et al. (2011) who expressed the need for a more fine-grained analysis of justifications. Accordingly, this work furthermore aims to develop an instrument that can—at least for some types of justifications—assess students' situation-specific justifications in a paper-and-pencil format. Even though work does exist that assesses students' progressions in scientific process skills by means of rubrics (Espinoza and Quarless 2010), that work has a broader emphasis than mere justifications. Currently, an assessment instrument that is suitable even for large sample sizes and that allows automated coding does not exist, although assessing student justifications in lab work is a future challenge for science education research (Henderson et al. 2018). This questionnaire can be a valuable tool for quantitative investigations about argumentation, for example, when analyzing how personal and situational factors influence justifications and how these influence learning.

Accordingly, we are interested in (a) which justifications are used by students to support or reject their own hypothesis on the basis of self-collected anomalous experimental data in lab work situations and how these justifications can be categorized, and (b) how students' use of different types of justifications can be empirically assessed in a qualified quantitative manner. The research questions are inextricably related: Of course, we can only *assess* students' justifications for and against a scientific hypothesis if we know *which* justifications students generate in a lab work setting.

Methods

The two research questions stated above call for different methodological approaches. Consequently, this paper reports on two sequentially conducted studies. First, we investigated research question (a) in Study 1 by means of interviews to obtain qualitative insights into the spectrum of possible justifications (as suggested by Chinn et al. 2011). Second, on the basis of the results of Study 1, we developed a quantitative instrument to assess selected justifications in order to answer research question (b) in Study 2. There is a strong justification for including both studies, despite the different methods used, in one paper: The studies have the same general aim of assessing justifications and reflecting on a consecutive and interdependent development process. Study 2 uses the results of Study 1 because original answers from students in the interviews were used to develop the questionnaire items. The questionnaire is, however, not simply an optimization of the interviews. The questionnaire focuses on only four of ten types of justifications identified in Study 1; the questionnaire aims to assess justifications in a time-efficient way by using multiple-choice items on a Likert-type scale. Importantly, we use the term “assessment” here in a value-free and unbiased way. We are not applying judgment regarding what counts as lesser or greater proficiency in justification. We are assessing the *nature of* and not the *quality or value of* students' justifications. For the sake of readability, we present these two studies in a logical rather than a chronological manner. In the following section, we first outline the general methods and design of both studies.

General Methods Used in both Studies

The Physical Contexts To answer the research questions, students had to engage in laboratory tasks. The goal of Study 1 was to elicit the broadest possible spectrum of justifications. It is

conceivable that the type of learning environment (hand-on vs. computer simulations), as well as the physics domain, might influence students' use of justifications. Accordingly, we chose to take into account both factors in the design of Study 1: First, we varied the type of the learning environment by the use of hands-on and computer-simulated experiments for one context. This was done because we wanted to consider the possible impact of the mode of collecting data (real vs. virtual) on justifications. Second, we varied the topic by using problems in mechanics and thermodynamics for the hands-on material. We did this to consider the possible impact of the topic on justifications.

Both topics met the following two conditions: (1) the context evoked a scientifically incorrect initial hypothesis for a large number of participants prior to the experiment, and (2) the hypothesis is easily testable for eighth- and ninth-grade students via quantitative experiments that produce anomalous data. Next, we outline both contexts and show that these requirements are fulfilled.

The Simple Pendulum Context Most students remember using a swing on a playground and hence have their own experiences with it. However, students usually have not conducted systematic observations with a swing. Hence, a large number of students believe that the time of oscillation of a simple pendulum depends on the mass of the pendulum. In Kanari and Millar's study (2004), this hypothesis was stated by 90% of the participants. Within the simplifications of the school curriculum, this is incorrect. A deeper and more general discussion shows that the mass of the pendulum does have a small influence on the period, but the effect is negligible in appropriate experimental settings (Nelson and Olsson 1986).

Hand-on materials and the computer simulation both allowed almost the same operations in the experimental space (Klahr 2000). In the hands-on experiment, students were given three masses, each weighing 50 g, that could be combined to create various weights. Adding mass did not affect the effective length of the pendulum. Hand-operated analog stopwatches were provided for measuring the pendulum's period (time). The angular displacement was measured on an angle meter and parallax errors had to be taken into account. The simulation experiment was based on the *pendulum lab* software of the *PhET* project of the University of Colorado, Boulder (PhET Interactive Simulations 2011). Various features of that experiment were altered or removed, including those features that allowed variation of string length and gravity. The version used for this study only allowed manipulation of mass and angular displacement. Measuring the period in the simulation experiment was accomplished with a digital stopwatch. The only difference between the hands-on and the computer-simulated experiments was the measurement uncertainty. The simulation calculates periods up to a tenth of a millisecond, whereas the stopwatch in the hands-on experiment has an uncertainty of 0.5 seconds due to reaction time. The demands on students' experimental competencies were low in both settings.

The Temperature in Solid Bodies Context For a vast majority of students, it is a common preconception to think that materials are able to "actively" warm something else (Albert 1978); "my winter jacket keeps me warm" is an example. The task was to explore the relationship between the insulation of a thermometer and the temperature it shows when the experiment is left in an environment with constant temperature. The experimental setup consisted of three foam cubes of different sizes and a digital thermometer. The foam cubes could be used to systematically vary the thickness of the sheathing, and thus the amount of insulation provided. Each foam cube had a hole drilled halfway into the cube, allowing the participants to place the thermometer in the center of the foam cubes. Like the simple pendulum experiment, this experiment is very easy to conduct.

Laboratory Task Procedures Both studies employed the same laboratory procedures. After a brief introduction to the contexts, participants were asked to state an initial hypothesis. To avoid irrelevant hypotheses (i.e., those that did not relate two variables, Hammann et al. 2006), the hypothesis space was artificially limited (Kanari and Millar 2004; Klahr 2000). Students had the choice of three hypotheses in the simple pendulum task: increasing the pendulum mass causes (a) an increase in the oscillation time, (b) no influence in the oscillation time, or (c) a decrease in the oscillation time. The temperature task also provided three hypotheses: increasing the thickness of the sheathing surrounding a thermometer causes (a) an increase in the temperature, (b) no influence in the temperature, or (c) a decrease in the temperature. Presenting three hypotheses in multiple-choice answer form can lead to “lucky guesses.” To circumvent this issue, students were asked to briefly explain their choice in an open text format immediately after choosing an initial hypothesis. Next, students tested their hypothesis experimentally with the available setup by collecting their own data. The experimental setting was constructed in such a way that the quality of the data allowed clear inferences. Although they had as much time available as needed, the experimental phase lasted 5 to 10 minutes.

Data Collection The process of data collection in both studies occurred right after the participants conducted the experiment. For Study 1, we immediately recorded semi-structured interviews with all the participants. Students were first asked whether they maintained or rejected their initial hypothesis. This was followed by a series of questions aimed at revealing the students’ justifications for their decisions: (1) “Do you keep your initial hypothesis?” (If initial hypothesis is kept go to question 3), (2) “Why do you reject your initial hypothesis?”, (3) “Why are you supporting/rejecting your hypothesis?”, (4) “How do you justify your decision?”, and (5) “How confident are you with your decision?” The goal of Study 1 was to generate the broadest possible spectrum of justifications. Hence, interviewers were free to ask more or different questions to elicit further justifications. This approach surely produces variance. This is the primary reason, among others, that we did not count frequencies across justifications. We will discuss this issue further later in the paper. The interviews lasted between 2 and 6 minutes ($M = 2$ minutes 57 seconds, $SD = 45$ seconds). All the participants were interviewed. Research question (b) aims to develop an instrument to assess students’ use of justification in a quantitative, valid, and reliable manner. Consequently, in Study 2, the interview was replaced by a questionnaire to assess students’ use of justifications in the process of data evaluation. Again, all the participants completed the questionnaire. The methodological approach and a detailed description of the development of the questionnaire are presented in the next sections.

Methodological Considerations Regarding the Quantitative Assessment of Justifications in Study 2

Students’ justifications of a hypothesis during lab work are embedded in an unobservable cognitive process. Therefore, it is necessary to induce a manifest behavior to quantitatively operationalize the different types of justifications. Much work in the field of argumentation consequently utilizes spoken language, such as studying group conversations (e.g., Riemeier et al. 2012), to look for evidence of implicit processes. This approach is justifiable considering the dialogical aspect of argumentation, yet it is also problematic in the context of our work, as it is possible that social desirability may create biased justifications during an interview (Nederhof 1985). For instance, a 13-year-old eighth grader interviewed face-to-face by an

unknown adult researcher regarding their justifications for or against an initial hypothesis might be afraid to say something he or she assumes might be seen as inappropriate in physics classes, for example, justifying a claim by referring to gut feelings. Utilizing written justifications (e.g., Kelly et al. 2007; Sandoval and Millwood 2005) might solve this problem, but it raises another issue: students may not be aware of certain justifications they use—such as intuition—and thus may not mention the justification explicitly in words.

Design of the Questionnaire

For the reasons outlined above, the most appropriate assessment method may be to anonymously ask students to rate their agreement with a list of justifications. This approach may both alleviate the social desirability bias of an interview structure (Nederhof 1985; Richman et al. 1999) and elicit implicit justifications. In addition, interviews and text analyses are time-demanding methods that are not suitable when studying structural relations between latent constructs because they usually demand a high sample size. Accordingly, research in this field must be studied in a quantitative as well as qualitative manner, which requires highly efficient and economical methods, such as self-administered questionnaires. Consequently, we chose to assess the justifications quantitatively by presenting students with a series of statements in a paper-and-pencil format and asking them to indicate to what extent these assertions applied to them in their own justification for supporting or rejecting their hypothesis. Nevertheless, we acknowledge here that “self-report assessment methods have been criticized as inauthentic, requiring respondents to construct or make explicit aspects of their epistemic cognition that many scholars believe to be largely tacit, complex, and contextual” (Sandoval et al. 2016, p. 21). We argue that this critique, though, is not justified here. In contrast to many of the studies this critique addresses, our instrument can only be employed within a certain situation, specifically, while conducting a scientific experiment. It would not be appropriate to use this instrument to assess epistemic cognition without any accompanying lab work. After conducting the experiment, the participants were asked to indicate the extent to which each statement applied to their justification. These statements, which we will subsequently call “items”, each correspond to one category of justification. For example, the item “When making this decision to keep or reject my assumption, I strongly considered the measurement data” refers to a category we name *data as evidence*. Consequently, we argue that students’ manifest rating of this item psychometrically measures the extent to which this category is used in their argument. Each item refers to only a single category.

Students rated the extent to which the items applied in their decision on a five-point Likert-type scale ranging from 0 (does not apply) to 4 (fully applies). The number of steps in Likert-type scales is extensively discussed in research (Cox III 1980). We decided to use a five-step Likert-type scale because we planned to analyze the data within the statistical framework of confirmatory factor analysis (CFA) using a maximum likelihood (ML) algorithm to estimate parameters (Brown 2006). There is sufficient evidence that a five-point scale fulfills the requirements regarding data quality using ML estimation (e.g., Beauducél and Herzberg 2006; Rhemtulla et al. 2012).

Development of the Questionnaire

The questionnaire was developed in three steps: (1) development of an item battery covering selected categories identified in Study 1, (2) evaluation of content validity, and (3) evaluation of psychometric quality. For the item battery, we used the coded interviews from Study 1 to establish a sufficient number of items. The fact that we developed items based on interview

data from Study 1 is already indicative of content validity because students' answers can be seen as verbalized forms of their justification, hence providing manifest elements of the latent construct "use of different types of justifications". However, because some of the passages taken from the interviews had to be rephrased during item development, the matching of items to the different justification types was ensured via expert rating, which we describe in the Results section. Finally, the content-validated set of items was analyzed to assess psychometric quality in terms of item difficulty, item variance, and discrimination. Furthermore, evidence is given below regarding the construct validity and the theoretical underlying factorial structure.

Data Analysis Procedures

Study 1 All the interviews from Study 1 were transcribed verbatim. We used an iterative grounded theory approach to derive categories from the transcripts. We began with the interviews from all the participants who worked on the simple pendulum task (both hands-on and simulation). Students' responses were broken into single artifacts whenever we observed that these artifacts contained justifications. Note that a student's complete answer may contain more than one artifact. These artifacts were then grouped by similarities into categories of justifications. Each category comprised a specific type of justification that could be used regardless of whether a student changed his or her initial hypothesis. Next, we analyzed the interviews from the students who worked with the temperature in solid bodies task in the same way to look for evidence of completeness, validity, and transferability of our categorization. To ensure the reliability of the coding process, two raters, who were trained on a small subset of transcripts using a coding manual, analyzed all the interviews. The raters decided for every justification identified in the transcripts whether a category does not apply (coded as 0), partly applies (1), or fully applies (2). Spearman's rank-order correlation coefficient was calculated to confirm inter-rater reliability, which is appropriate for ordered categorical judgments (Gwet 2010). The average correlations are $\rho = .80$ (simple pendulum) and $\rho = .61$ (temperature in solid bodies). In addition, the percentage agreement was analyzed, resulting in averages of 91.4% (simple pendulum) and 85.4% (temperature in solid bodies). Detailed results for inter-rater reliabilities for each category are provided in Table 1.

Study 2 All the data from Study 2 were extracted into a tabular format. Responses from the Likert-scaled items were coded from 0 (does not apply) to 4 (fully applies). Mardia's test (1970) indicated that the data violates the assumption of multivariate normal distribution (skew $S = 233.3$, $p_S < .001$, kurtosis $K = 3.96$, $p_K < .001$; criterion: $S < 2$, $K < 7$; see West et al. 1996). Accordingly, we used robust ML estimators in the CFA, which are seen as unbiased when using ordinal indicators with five or more steps (Pui-Wa and Qiong 2012; Rhemtulla et al. 2012) and which can further handle missing values by employing full-information ML methods (Enders and Bandalos 2001). All the models were calculated within the statistical framework R (R Core Team 2014) using the software package lavaan (Rosseeel 2012). The metric of the latent variable was scaled by fixing the variances of the latent factors to 1 (Brown 2006).

Samples

In line with research question (a), we aimed to generate a broad spectrum of possible justifications in Study 1. Following a "sampling for variation" approach (Morse and

Table 1 Inter-rater reliabilities determined by Spearman's rank-order correlation coefficient and percentage agreement

Category	Pendulum		Temperature	
	Percentage agreement	ρ	Percentage agreement	ρ
Appeal to an authority	96.6%	1	95.8%	.59
Data as evidence	82.8%	.54	47.9%	.34
Experimental competence (technical/skills)	96.6%	1	97.9%	.84
Experimental competence (self-concept)	100.0%	1	97.9%	a
Ignorance	96.6%	.85	95.8%	.75
Intuition	93.1%	.71	91.7%	.46
Measurement uncertainties (explicit)	93.1%	.85	93.8%	.88
Measurement uncertainties (implicit)	82.8%	.70	62.5%	.58
Suitability of experimental setup	96.6%	.83	95.8%	.82
Use of theoretical concepts	75.9%	.60	75.0%	.81

All correlations are significant ($p < .05$)

^aNo correlation could be computed because one rater did not observe this category

Niehaus 2009, p. 65), we recruited $n = 129$ students from grades 8 through 10 from four different high schools in an urban area of Germany. The students' ages ranged from 12 to 18 years ($M = 14.6$ years, $SD = 1.1$ years). The participants were randomly assigned to the experimental tasks ($n = 38$ to the simple pendulum hands-on task, $n = 41$ to the simple pendulum simulation task, and $n = 48$ to the temperature in solid bodies task). Study 2 included 154 students from grades 8 through 10 (age $M = 14.4$, $SD = 1.1$; 45.2% girls) attending a single high school in an urban area of Germany. They were randomly assigned to either the hands-on or computer-simulated simple pendulum context. All the students had sufficient experience in experimentation (e.g., could read a stopwatch). To avoid the influence of prior content knowledge, we asked the participants' teachers not to address the relevant content during their regular science classes in the weeks before the study (e.g., regarding the relationship between the mass of the bob and the time of swing). As expected, the vast majority of students stated an incorrect initial hypothesis (88% in the pendulum context and 82% in the temperature context). Students who did not choose an incorrect initial hypothesis were not included in the analysis in either study, as it could be assumed that those students did not conceive the data as anomalous. In contrast, we assume that those students with incorrect initial hypotheses saw a discrepancy between their hypotheses and their own collected experimental data.

Results

Study 1: Which Justifications Are Used and How Can These Be Categorized?

The analysis of all the interviews from both contexts led to ten categories of justifications: *appeal to an authority*, *data as evidence*, *experimental competence (technical/ skills)*, *experimental competence (self-concept)*, *ignorance*, *intuition*, *measurement uncertainties (explicit)*, *measurement uncertainties (implicit)*, *suitability of the experimental setup*, and *use of theoretical concepts*. Table 2 describes these categories in general and provides example statements from the simple pendulum task. In the following, we give two examples from the interviews to demonstrate how the participants' answers were assigned to categories.

Table 2 Types of justifications for supporting or rejecting an initial hypothesis after confrontation with self-collected, quantitative, anomalous experimental data

Category	Description	Example
<i>Appeal to an authority</i>	Statements that explicitly refer to experts or to the need for verification from experts; participants may change their initial hypotheses but express skepticism due to lack of assurance from an expert (e.g., teacher)	<i>“The time of swinging always stayed equal. [...] I think that you tried these experiments before, so I can expect that it works. The computer uses formulas to calculate the time.”</i> <i>“Somebody else should tell me that it [the new hypothesis] is correct.”</i>
<i>Data as evidence</i>	Statements referring explicitly or implicitly to experimentally derived data (including observations) as evidence	Explicit: <i>“During my observations, it stayed the same, no matter the weight. I tried once with 0.1 and once with 2 kg and it was always the same. Always 2.958 or something like that.”</i> Implicit: <i>“It has been proven by the experiment that the time always stayed the same.”</i>
<i>Experimental competence (technical/-skills)</i>	Statements that address one’s own competency in doing experimental work with respect to technical issues	<i>“It could be that I have somehow measured wrong, so once too early, once too late.”</i>
<i>Experimental competence (self-concept)</i>	Statements that address one’s own competency in doing experimental work with respect to self-efficacy and self-concept	<i>“I do not know exactly. There is still something in the back of my mind. It might be that I did something wrong. I was always bad in physics and experimentation.”</i>
<i>Ignorance</i>	Statements that mention the measurement results but include a hypothesis that is contradictory to these results	<i>“I did not measure any differences [in the time of swinging] when I put the pieces of mass on the pendulum.”</i> But afterwards the participant insisted on a contradictory hypothesis: <i>“The bigger the pendulum mass, the smaller the time of swinging.”</i>
<i>Intuition</i>	Statements that can be attributed to an emotionally driven, intuitive view of the anomalous data	<i>“I seem to remember that the time [of swinging] stayed equal. But my gut feeling says that if the pendulum’s mass is bigger, it swings longer!”</i>
<i>Measurement uncertainties (explicit)</i>	Statements that directly refer to measurement uncertainties and show awareness of uncertainty in data	<i>“There may somehow be measurement errors.”</i>
<i>Measurement uncertainties (implicit)</i>	Statements that, through misleading interpretations of the data, show naïve conceptions about measurement uncertainties; in contrast to the category above, these justifications do not explicitly refer to uncertainty	<i>“I changed the weight [of the pendulum] and the stopwatch said 2.94 and then 2.96 seconds. So, the time of swinging becomes longer.”</i>
<i>Suitability of experimental setup</i>	Statements questioning the overall suitability of the setup, e.g., concerning control of variables or constraints or questioning the assumed precision/accuracy of the experimental setup or simulation	<i>“I was quite sure of my initial hypothesis, but then I got something totally different. The computer told me that it is always the same time—it does not make sense to me. [...] I should trust the computer, actually. But I’m quite unsure, because other influences like wind or friction are not considered in here.”</i>
<i>Use of theoretical concepts</i>	Statements that try to explain the results with a (typically wrong) scientific theory or ad hoc explanation in order to avoid confrontation with the anomalous data	<i>“The more weight I have on the pendulum the shorter the time of swinging. This is because the weight tightens the thread and the pendulum is much faster then.”</i>

Interview 1 (with an 8th-grade student, 14 years old, simple pendulum task, hands-on experiment, wrong initial hypothesis):

Interviewer: You just did the experiment. Do you keep your initial hypothesis?

Student 1: Well, yeah!

Interviewer: Why?

Student 1: I found out that the heavier the pendulum, the longer it takes [the time of oscillation]. The difference wasn't much, just a few hundredths of a second. Once I had 2.97 s when I had 200 g on the pendulum. Then I had way less [mass] but only had like 2.96 s. The difference is just minimal.

We identified two types of justifications in this interview. (1) The student referred to measurement *data as evidence* to justify the decision (third and fourth sentence). (2) The student expresses that the differences between two measurements are small (second and last sentence) and concludes that there is a main effect (even though the effect is small). This hypothesis justification indicates a lack of knowledge of measurement uncertainties. Because the student did not *explicitly* refer to measurement uncertainties, we name this category *measurement uncertainties (implicit)*.

Interview 2 (with an 8th-grade student, 13 years old, simple pendulum task, simulation experiment, wrong initial hypothesis):

Interviewer: You just did the experiment. Do you keep your initial hypothesis?

Student 2: No.

Interviewer: Why?

Student 2: Because I thought the mass isn't that important. I assumed that the greater the mass, the longer the time of swing. Because the air drag is bigger then. But that doesn't have anything to do with it, right? In my observations, it was the same [the time of oscillation] no matter which weight I used. I measured once with 0.1 kg and once with 2 kg and it always stayed the same. Always 2.958 or something.

Interviewer: How confident are you with your decision?

Student 2: I'm dead certain!

Interviewer: Why?

Student 2: Because I observed something totally different when I conducted the experiment. The computer has its reasons for that.

Interviewer: Can you explain why?

Student 2: Yeah, that simulation was programmed at the university. It has to be correct.

Interviewer: What is your new hypothesis then?

Student 2: The time of swing stays the same if you change the mass. I think it's more about the displacement than about the mass.

This student's interview displays three types of justifications. (1) The student refers explicitly to measurement *data as evidence*. (2) The student justifies the decision by *appealing to an authority* by mentioning that the software was programmed at a university and thus must be correct. (3) In the final sentence, the student uses a new *theoretical concept* (introducing angular displacement as a possible factor that has an effect on the time of oscillation) to justify a "new" hypothesis.

Study 2: How Can the Use of Different Types of Justifications Be Empirically Assessed?

Selection of Categories for Test Development Research question (b) aims to develop an instrument to empirically assess the use of different justifications in lab work learning situations. Because Study 1 led to ten different categories of justifications—which is quite a great deal in terms of test development—we selected four categories for operationalization in Study 2: *intuition*, *appeal to an authority*, *measurement uncertainties (explicit)*, and *data as evidence*. This selection, which we will elaborate on below, was based on the general relevance of the justification types for learning science with respect to the literature. *Intuition* was included because students gave non-rational justifications in the interviews as a matter of course. The use of this type of justification is particularly relevant for science education research, as it relates to both hot vs. cold conceptual changes (see above) and the “wish” of educators to teach students to justify claims on a rational basis such as data obtained from experiments. Further, intuition is known to be an important factor for learning science, but it has received little attention in science education research (Fensham and Marton 1992). Investigating the use of intuition in lab work is particularly relevant, as it is well known from other disciplines that people tend to rely on intuition in statistical decision-making (Kahneman and Tversky 1973). *Appeal to an authority* was chosen because students’ interviews indicated that this type of justification was used to signify a wide range of authorities, including teachers, information sources, and experiments. Additionally, it is highly relevant to know the extent to which expertise, which is already integrated into the experiment because an expert put it together for the students, influences cognition during processes of data evaluation and experimental observations (Hug and McNeill 2008). Finally, given today’s easy access to information using digital technology, it is important to know the extent to which students trust their sources when justifying claims. *Measurement uncertainties (explicit)* was used because students’ answers revealed an awareness of, albeit little knowledge of, measurement uncertainties (Priemer and Hellwig 2018). Little work in the context of argumentation examines the influence of measurement uncertainty in data to justify a claim, despite the fact that evaluating quantitative data in order to draw conclusions is not possible without estimating the uncertainty (e.g., Lee et al. 2014). For obvious reasons, the category *data as evidence* was operationalized because the justification of claims on the basis of measurement data used as evidence is at the core of science and is addressed in science standards (e.g., NGSS Lead States 2013).

Item Authoring and Content Validity The developed items were all based on the interview answers coded in Study 1. Where possible, we used students’ original statements—sometimes even citing complete sentences—to develop the items. Additionally, for the categories *intuition* and *data as evidence*, we were able to draw on established instruments, such as the Rational-Experiential Inventory (Epstein et al. 1996; Keller et al. 2000) and a corresponding German inventory (PID, Betsch 2004). This approach allowed us to develop an item battery of 88 items: 25 items in the category *intuition* (example item: “I relied heavily on my gut feelings when I made conclusions about the experiment,” #int03, see Online Resource 1), 16 items in the category *appeal to an authority* (“When making my decision, I took into account that experts put the experiment together and it should function properly,” #aut05), 21 items in the category *measurement uncertainties (explicit)* (“My conclusions took into account that the experiment contains measurement inaccuracies,” #mu04), and 26 items in the category *data as evidence* (“I chose to keep or reject my assumption based on the measurement data,” #evi05). To further ensure content validity, eight graduate students (26–34 years old) from different domains (three in physics, two in chemistry, one

in biology, one in English, and one in arts) were asked to judge item texts regarding the category of justification they address. The broad academic backgrounds ensured that the content validity could be extended beyond the domain of physics. All the participants were either close to completing or already attained a master's degree in their respective fields and were pursuing a doctoral degree in science education. The experts were first presented with a detailed description of the justification types. On the basis of these descriptions, we asked the experts to assign each item to one of the four categories with an additional option of "none of the above". We established the following criterion for content validity: If seven of eight experts assigned an item to the intended category, then we accepted this item as valid (percentage agreement on item level of 75% (Gwet 2010)). Among the 88 items, 63 items matched this criterion: 18 for the category *intuition*, 13 for the category *measurement uncertainties (explicit)*, 12 for the category *appeal to an authority*, and 20 for the category *data as evidence*. We must note that for six items in the category *appeal to an authority*, only six (instead of seven) of the eight experts agreed on the classification; however, we chose to include these items in the questionnaire. In this case, two experts claimed that the text in these six items did not explicitly refer to "expert knowledge" and so could not unequivocally be assigned to *appeal to an authority*. To correct this, we rephrased those six items and put stronger emphasis on expert knowledge. The coding reliability corrected for chance agreement for all 57 items (rephrased items were excluded) and had a Fleiss' kappa value of .90, which can be interpreted as excellent (Gwet 2010). Six inadequate phrasings of items were identified and revised. Although these six items were not re-rated by the experts, we argue in favor of the content validity of these items because the deficiencies were obvious upon comparison with valid items. A set of 63 content-valid items were subjected to an evaluation of psychometric quality.

Analysis of Psychometric Quality To further reduce the 63-item battery to a reasonable number of items, we initially used two criteria to select the items with the highest psychometric quality: (1) Items with extreme item difficulty were excluded. We want to denote that item difficulty is a technical term that does not imply that the instrument assesses an underlying construct of ability or skill (Kline 2000). (2) The discrimination in terms of part-whole corrected item-total correlation had to be sufficiently large ($> .4$, Weiner et al. 2012). Item selection based solely on discrimination carries the risk of low variance, that is, the risk that the resulting scale will not be able to differentiate sufficiently across a wide range of test scores, as variances in item difficulty naturally reduce discrimination. Lienert's selection index (Lienert and Raatz 1998) tries to overcome these issues by combining discrimination, variance, and item difficulty into an integrated measure. Consequently, we added a third criterion to the item selection: (3) The items with the largest selection indices were selected from the remaining set of items, not exceeding eight items per category of justification. This procedure led to the establishment of a set of 31 items. The item difficulty of the remaining items falls in the medium range, but the items overlap at certain intervals (*intuition*: $25.8 < P_i < 46.5$; *appeal to an authority* $48.3 < P_i < 58.5$; *measurement uncertainties (explicit)* $43.3 < P_i < 51.3$; *data as evidence* $60.5 < P_i < 73.8$). The discriminatory index is consistent with the criterion $r > .4$, while for three of the four scales, most items reached higher values (*intuition* $.52 \leq r_i \leq .76$; *appeal to an authority* $.68 \leq r_i \leq .78$; *measurement uncertainties (explicit)* $.40 \leq r_i \leq .53$; *data as evidence* $.53 \leq r_i \leq .70$).

Because "prior theory regarding the structure of the data" (Henson and Roberts 2006, p. 409) exists, we conducted a CFA to evaluate the underlying four-factor structure of our questionnaire. In addition, CFA estimates the discriminant and convergent validity and allows testing of competing models (for a more in-depth review of the features of CFA, see Brown 2006). To estimate the factorial validity, we defined four competing models. Models 1–3 include four

factors that reflect the four categories of justification. Model 1 contains all 31 items, while Models 2 and 3 use a reduced version of Model 1 with five items per category (items with the highest CFA-based factor loadings). Models 2 and 3 differ in that Model 2 allows for covariance of the factors, while Model 3 has an orthogonal factor structure. Model 4 represents the simplest factor structure wherein all items load on one latent factor (which could be interpreted as students’ general tendency towards agreement with all items in the same manner). All evaluations of the model fits were accomplished with regard to accepted standards of interpretation of fit indices (Hu and Bentler 1999). The model fit analysis adhered to the following rules: For models exceeding 100 degrees of freedom, the χ^2 test was not interpreted because the assumption of an “exact model fit” is untenable (MacCallum et al. 1996, p. 144). Instead of the χ^2 test, the *comparative fit index* (CFI), *root mean square error of approximation* (RMSEA), and *standardized root mean square residual* (SRMR) were interpreted. A CFI of $> .90$ is widely seen as the cutoff criterion (MacCallum et al. 1996; West et al. 2012), whereas $< .08$ is the cutoff for the RMSEA, and $< .08$ is the cutoff for acceptable to good fits for the SRMR (Brown 2006). The goodness-of-fit indices for all the models are reported in Table 3.

Model 1 shows a poor fit, as the CFI is too low. Model 2, which includes only a small subset of items, indicates a reasonably good model fit, as all goodness-of-fit indices meet their criteria (CFI $> .90$, RMSEA $< .08$, SRMR $< .08$). The competing Model 3 is noticeably worse than Model 2 because all the fit indices are lower than in Model 1. Model 4, a one-factor model, is discarded because not all the fit indices meet the cutoff criteria. Hence, Model 2 shows the best fit to the data and was used for further analysis. The standardized factor loadings, which can be interpreted as discrimination indices (Brown 2006), all show reasonably high values ($\bar{\lambda} = .71$, $SD = .10$, $\lambda_{\min} = .50$, $\lambda_{\max} = .84$). The correlation among latent factor structures is only significant for three relations (with medium or low effect sizes): the correlation between the factors *intuition* and *data as evidence* ($r_{\text{lat}} = -.50$, $p < .001$), between *appeal to an authority* and *measurement uncertainties (explicit)* ($r_{\text{lat}} = -.34$, $p = .01$), and between *appeal to an authority* and *data as evidence* ($r_{\text{lat}} = .23$, $p = .04$). The reliability of the scales can be calculated directly from the CFA model in terms of the proportion of true variance to total variance of the measurement (Brown 2006). The CFA-based reliabilities are .86 for *intuition*, .90 for *appeal to an authority*, .72 for *measurement uncertainties (explicit)*, and .82 for *data as evidence*.

Discussion

The aim of Study 1 was to identify different types of justifications that students give for hypotheses in physics, when faced with quantitative anomalous data obtained from their own

Table 3 Goodness-of-fit indices for Models 1 through 4

#	Model	χ^2	<i>df</i>	<i>p</i> value	$\chi^2/$ <i>df</i>	CFI	RMSEA [90% CI]	$P_{\text{RMSEA} < .05}$	SRMR
#1	4-factor model, all 31 items	599.5	428	.00	1.40	.88	.05 [.04; .06]	.23	.08
#2	4-factor model, 5 items per category	216.4	164	.00	1.32	.94	.05 [.03; .06]	.54	.07
#3	4-factor model, orthogonal, 5 items per category	258.6	170	.00	1.52	.90	.06 [.05; .08]	.09	.13
#4	1-factor model, 5 items per category	818.4	170	.00	4.81	.29	.17 [.16; .18]	.00	.19

experiments. We found ten different categories of justifications. The existence of some of the categories is supported by the results of other studies, including the use of intuition (Fensham and Marton 1992; Sandoval and Çam 2011), ignoring data (Chinn and Brewer 1998; Sandoval and Çam 2011; Kelly et al. 2001), referring to theoretical concepts instead of analyzing own data (Sandoval and Morrison 2003; Walton 1996), appealing to an authority (Hug and McNeill 2008; Sandoval and Çam 2011; Walton 1996), and the use of evidence (Chinn and Brewer 2001; Jeong and Songer 2008; Khine 2012; McNeill and Berland 2015; Roberts and Gott 2010; Scheiner 2004; Walton 1996). Other categories are novel to the analysis of justifications, such as those referring to experimental competences, measurement uncertainties, and the suitability of the experimental setup. These findings show that our initial assumption—that the existing classifications are not connected to learning environments with students' own experimental data—was correct. This points to an important difference between our study and that of Chinn and Brewer (1998): While we investigated student-generated experimental data, they used entire theories (with plausible predefined initial hypotheses) and predefined data sources. We argue that using predetermined data eliminates aspects of students' interaction between the hypothesis and the experiment space (Klahr 2000). The fact that we found both previously known and new categories shows that argumentation is influenced by the situation in which it occurs and that it adheres to general strategies (Kind 2013).

The identified justification types indicate that there are rational (e.g., *data as evidence*) and non-rational (e.g., *intuition*) justifications. Hence, we argue that non-rationality in argumentation in science instruction should not be overlooked (Sinatra 2005). It may, for example, be the case that students make non-rational decisions when they do not have enough information at hand to make informed choices. This is in line with Petty and Cacioppo (1986), who state in their elaboration of likelihood model of persuasion (ELM), that an inability to process information can lead to peripheral (non-rational) choices. In our study, inability to process the information appears as students' difficulties in manipulating and evaluating the data. These problems may lead to the use of the justification types *ignorance*, *measurement uncertainties (implicit)*, and *use of theoretical concepts*. The latter of which may be based on conceptions grounded in students' every-day lives.

It may at first seem appropriate to simply use the categorization developed by Chinn and Brewer (1998), as our aim—like theirs—is to systematically describe students' responses to anomalous data. However, in Chinn and Brewer's (1998) taxonomy, the assignment of an artifact to a certain response type is dependent upon whether the student supported or rejected their initial hypothesis. We argue that this is a restriction. For instance, this is clearly the case for our category *appeal to an authority* in which the assertion "*Someone has to tell me that this is true!*" (see Table 2) can occur as a justification, either for or against an initial hypothesis. Thus, in contrast to Chinn and Brewer's (1998) taxonomy, our categories of justifications are not segregated based on whether a student chooses to keep or reject their initial hypothesis.

It is worth noting that another study by Lin (2007), which also used laboratory experiments, identified 17 categories of justifications. Although all 17 categories can be matched to our categorization, the allocation is rather problematic, as it includes some overlap. For example, one of Lin's (2007) categories, *accept anomalous data but do not know why*, can be allocated to two of our categories: *data as evidence* and *intuition*. Furthermore, Lin (2007) suggests that this justification (*accept anomalous data but do not know why*) does not lead a participant to a conceptual change, even though anomalous data are accepted. We doubt that this is sufficiently proven. Accepting data, even without knowing why, can still involve a conceptual change. Likewise, problems arise here again, as Lin (2007) used Chinn and Brewer's categories (1998),

and in so doing, distinguished between the justifications based on whether students supported or rejected their initial hypothesis. Hence, it can be concluded that the classification systems in both our study and Lin's (2007)—and by extension, Chinn and Brewer's (1998) study—were created differently and serve different functions.

The results of this study are also particularly relevant for practitioners. Our proposed categorization of justifications given by students allows science teachers to anticipate the possible range of justifications students might generate in the context of lab work. Thus, our work provides teachers with relevant input regarding what to expect from students' answers so that they can prepare themselves to respond to the students' justifications in an appropriate way. For example, if students refer to gut feelings or ignore data completely when justifying a claim, teachers can provide help by supporting students with prepared worksheets that focus on the evaluation of the evidence at hand. This might include a discussion of measurement uncertainties when estimating the quality of the data. Thus, awareness of the fact that students will not always use justifications that are favored by science educators (e.g., from the category *intuition*) can help teachers to address this in the classroom. More detailed implications of the results for practice are described elsewhere (Ludwig and Priemer 2017).

In summary, three new aspects characterize our Study 1 results: (a) The use of self-collected experimental data are now included in the assessment of justifications. (b) Contexts with relations to every-day life experiences are now considered. (c) The types of justifications do not depend on a student's decision to maintain or change an initial hypothesis or concept. The latter is valuable because conceptual change is hard to reach and to assess (Posner et al. 1982).

Of course, our categorization has limitations. Given that the categories of justifications found in the simple pendulum task could also be found in the temperature in solid bodies task, we conclude that our categorization is discerning and comprehensive within the scope of the methodology used. However, as noted above, justifications can be context-dependent (Chinn et al. 2011; Sandoval 2012). By addressing two contexts in our study, we made sure that the results do not depend on a single topic. Hence, we provide a good starting point for further research. It is conceivable that most categories might appear in situations outside of physics as well because they are relatively independent from the content: This might account for the categories *data as evidence*, *intuition*, *appeal to an authority*, *ignorance*, *suitability of the experimental setup*, *experimental competence (technical/skills and self-concept)*, and *use of theoretical concepts*. The categories *measurement uncertainties (explicit and implicit)* are probably mostly encountered in lab work situations in which the uncertainty of measurement plays a major role, which is more often the case in the domain of physics than in other classes. Scholars and practitioners should be aware of this constraint.

Both physics contexts of our studies have further characteristics that limit their generalizability. These are related to students' every-day life experiences and their proneness to non-scientific preconceptions. Further, our participants had little prior knowledge in these contexts, which may influence their use of justifications, for example, when they refer to known theories. While conducting the experiments to generate their own quantitative data, the students collected evidence of varying quality and quantity (for example, number of repetitions and precision of measurements) and documented their results differently. This led to variation in the resources students had at hand to recapitulate their experimental work when giving justifications. It remains an open question whether our justification types are valid in settings in which students are not required to interpret anomalous data. Furthermore, we do not know if younger or older students would use the same justifications. We emphasize that we did not determine the frequency of the use of the different justification types for the following reasons: First, the

common practice of quantitative analysis of qualitative data is commonly criticized (Hammer and Berland 2014). Second, to compare the frequencies across categories would have required us to ensure that the interviews elicited all justifications a student might have in mind. This was not the aim of the interview questions. Finally, the students' answers might have been biased, as described earlier, by social desirability. This is especially conceivable with categories such as *intuition*, which students might see as inappropriate in science classrooms and therefore might not mention, even if such categories did play a role in their decisions.

To eliminate some of these limitations and to enable further research into justifications of hypotheses, Study 2 aimed to develop an instrument to assess students' use of justifications in a valid and quantitative manner. The four selected categories (*data as evidence*, *measurement uncertainties (explicit)*, *intuition*, and *appeal to an authority*) are of general interest in lab work settings because they reflect known non-rational justifications that are also found in other studies (such as using intuition and referring to experts) and because they focus on the evaluation of the collected data (evidence and evaluating the quality of data with respect to measurement uncertainties). The questionnaire can be administered in 5–10 minutes and is thus especially suitable for medium- to large-scale assessments. The target group is eighth- and ninth-grade students. We note that the questionnaire does not assess students' generally preferred justifications for hypotheses. The results obtained from this questionnaire are also highly contextual and are directly related to the specific laboratory task. However, this is a "feature" and not a disadvantage, as it is in line with research that demands fine-grained measures that meet the requirement of the contextual character of learning processes (Chinn et al. 2011).

Defending the validity of the instrument is crucial in the development of this new tool to assess students' use of justifications. The results of the CFA confirm the claimed underlying four-factor structure of the questionnaire. All the items yielded consistently high factor loadings ($> .6$). This indicates good convergent validity of the questionnaire (Brown 2006). The absence of a strong correlation between latent factors is essential for gaining evidence for divergent validity. Brown (2006) sets .8 as a cutoff value. In our work, only three (of six possible) structural correlations between factors yielded significance, but these were moderate in size. This speaks to the fact that all four underlying categories of justification can be measured independently. Model-based reliability estimators consistently return high results. Accordingly, we argue that the questionnaire measures the use of justification in a reliable manner. Thus, we gained evidence for the quality of the questionnaire in terms of content validity (expert rating), factorial validity (CFA), discriminant validity (correlations between factors), convergent validity (high factor loadings of indicators), and reliability (model-based reliability estimators). Hence, we argue that the questionnaire is an appropriate method of analyzing students' use of justifications generated from data and experimental observations in a given laboratory situation.

The final questionnaire is available as [supplementary material](#) accompanying the online version of this article (Online Resource 1). Note that an English translation is provided here; however, the questionnaire was developed and distributed in German. Thus, the authors caution that the textual validity of the translated items has not been explicitly investigated. Nevertheless, the questionnaire and the data we have collected concerning it, provide a good basis for further research. Of course, the psychometric evaluation of the instrument also has limitations. Due to the exclusion of items in the development of the test, the four justification categories may be underrepresented in the questionnaire. Moreover, the questionnaire may not be able to sufficiently differentiate between persons in the very extreme lower and upper ends of the scale due to the lack of items in these ranges. While the chosen model, Model 2, shows a

satisfactory fit to the empirical data, other items representing the four justification categories might have led to another model and hence possibly a better model fit. Finally, we used only a limited number of competing models, specifically only one- and four-factor models. We thus do not know whether another factorial structure may have a better fit than Model 2. Again, we want to stress that we decided against reporting the means of the newly developed scales due to a possible research bias that might occur when scale analysis, item selection, and hypothesis testing are carried out on the same sample (Kline 2000).

The questionnaire in its present form is a valuable tool for assessing justifications frequently used by students. For example, researchers can now study highly relevant issues related to the epistemic process of justifying a scientific claim: How do situational factors such as the use of computer-simulated experiments versus hands-on material or the quality of data gained from different setups influence students' use of different types of justifications? How do personal factors, such as the ability to evaluate data or domain-specific knowledge, influence how students justify hypotheses? Will a highly motivated student automatically justify claims on the basis of measurement data as evidence? Will the use of justifications vary with age? Is it possible to foster students' use of data as evidence by means of explicitly designed learning environments? Knowing what influences the use of justifications, how this develops over time, and how we can promote the use of rational justifications rather than non-rational justifications is only one side of the same coin: Science educators also have to investigate how the use of different types of justifications may affect learning outcomes in the science lab. Because it is conceivable that, for example, the use of data as evidence or the evaluation of uncertainties might lead to better learning outcomes. These hypotheses can now be empirically investigated. Furthermore, justifications are components of arguments that are often built to persuade—whether oneself, a classmate, or the scientific community—therefore, it seems especially important to investigate the relationship between persuasion and the use of different justifications. Some of these questions are currently undergoing further research, in which we are applying our questionnaire to two large-scale studies with more than 1500 high school students participating in lab work courses. Employing methods of latent variable modeling, our first analysis shows that, for example., students prefer to use data as evidence if they have a certain ability to evaluate data. Again, if students use data as evidence in their justification, they are more likely to state a correct hypothesis after experimentation, while relying on intuition leads to a less permanent decision (Ludwig 2017). We could also demonstrate that the learning environment—real vs. virtual—does not seem to influence students' use of different justifications (Ludwig and Priemer 2013). These results point to the fact that it is now possible to investigate the process of stating a scientific hypothesis based on experimentally derived data at a fine-grained level.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Abi-El-Mona, I., & Abd-El-Khalick, F. (2011). Perceptions of the nature and 'goodness' of argument among college students, science teachers, and scientists. *International Journal of Science Education*, 33(4), 573–605.
- Albert, E. (1978). Development of the concept of heat in children. *Science Education*, 62(3), 389–399.

- Anderson, R. D. (2007). Inquiry as an organizing theme for science curricula. In S. K. Abell & N. G. Ledermann (Eds.), *Handbook of research on science education* (pp. 807–830). London: Lawrence Erlbaum.
- Asterhan, C. S. C., & Schwarz, B. B. (2009). Argumentation and explanation in conceptual change: indications from protocol analyses of peer-to-peer dialog. *Cognitive Science*, 33(3), 374–400.
- Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(2), 186–203.
- Betsch, C. (2004). Präferenz für intuition und deliberation (PID) (Preference for intuition and deliberation). *Zeitschrift Für Differentielle Und Diagnostische Psychologie*, 25(4), 179–197.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.
- Brown, D. E., & Hammer, D. (2008). Conceptual change in physics. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 127–154). New York, NY: Routledge.
- Chan, C., Burtis, J., & Bereiter, C. (1997). Knowledge building as a mediator of conflict in conceptual change. *Cognition and Instruction*, 15(1), 1–40.
- Chinn, C. A., & Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching*, 35(6), 623–654.
- Chinn, C. A., & Brewer, W. F. (2001). Models of data: a theory of how people evaluate data. *Cognition and Instruction*, 19(3), 323–393.
- Chinn, C. A., & Malhotra, B. A. (2002). Children's responses to anomalous scientific data: How is conceptual change impeded? *Journal of Educational Psychology*, 94(2), 327–343.
- Chinn, C. A., Buckland, L. A., & Samarapungavan, A. (2011). Expanding the dimensions of epistemic cognition: arguments from philosophy and psychology. *Educational Psychologist*, 46(3), 141–167.
- Cox, E. P. C., III. (1980). The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research*, 17(4), 407–422.
- Delen, I., & Krajcik, J. (2015). What do students' explanations look like when they use second-hand data? *International Journal of Science Education*, 37(12), 1953–1973.
- Department for Education. (2015). National curriculum in England: science programmes of study. Retrieved from <https://www.gov.uk/government/publications/national-curriculum-in-england-science-programmes-of-study/national-curriculum-in-england-science-programmes-of-study>. Accessed 2 June 2017.
- Dole, J. A., & Sinatra, G. M. (1998). Reconceptualizing change in the cognitive construction of knowledge. *Educational Psychologist*, 33(2), 109–128.
- Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education*, 84(3), 287–312.
- Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Studies in Science Education*, 38(1), 39–72.
- Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 8(3), 430–457.
- Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology*, 71(2), 390–405.
- Espinoza, F., & Quarless, D. (2010). An inquiry-based contextual approach as the primary mode of learning science with microcomputer-based laboratory technology. *Journal of Educational Technology Systems*, 38(4), 407–426.
- Fensham, P. P., & Marton, P. F. (1992). What has happened to intuition in science education? *Research in Science Education*, 22(1), 114–122.
- Ford, M. J. (2012). A dialogic account of sense-making in scientific argumentation and reasoning. *Cognition and Instruction*, 30(3), 207–245.
- Ganser, M., & Hammann, M. (2009). Hypothesen verändern können. Aufgaben zum Umgang mit unerwarteten Daten im Kontext historischer Experimente (Change hypotheses. Learn how to deal with anomalous data in the context of historical experiments). *Praxis der Naturwissenschaften, Biologie in der Schule*, 58(3), 39–43.
- Gregoire, M. (2003). Is it a challenge or a threat? A dual-process model of teachers' cognition and appraisal processes during conceptual change. *Educational Psychology Review*, 15(2), 147–179.
- Gwet, K. L. (2010). *Handbook of inter-rater reliability*. Gaithersburg: Advanced Analytics, LLC.
- Hammann, M., Phan, T. H., Ehmer, M., & Bayrhuber, H. (2006). Fehlerfrei Experimentieren (Flawless experimentation). *Der Mathematische Und Naturwissenschaftliche Unterricht*, 59(5), 292–299.
- Hammer, D., & Berland, L. K. (2014). Confusing claims for data: a critique of common practices for presenting qualitative research on learning. *Journal of the Learning Sciences*, 23(1), 37–46.
- Hammer, D., & Elby, A. (2002). On the form of personal epistemology. In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: the psychology of beliefs about knowledge and knowing* (pp. 169–190). Mahwah: Lawrence Erlbaum Associates.

- Henderson, J. B., McNeill, K. L., González-Howard, M., Close, K., & Evans, M. (2018). Key challenges and future directions for educational research on scientific argumentation. *Journal of Research in Science Teaching*, 55(1), 5–18.
- Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research common errors and some comment on improved practice. *Educational and Psychological Measurement*, 66(3), 393–416.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55.
- Hug, B., & McNeill, K. L. (2008). Use of first-hand and second-hand data in science: does data type influence classroom conversations? *International Journal of Science Education*, 30(13), 1725–1751.
- Jeong, H., & Songer, N. B. (2008). Understanding scientific evidence and the data collection process: explorations of why, who, when, what, and how. In C. L. Petroselli (Ed.), *Science education issues and developments*. New York: Nova Science Publishers.
- Jeong, H., Songer, N. B., & Lee, S.-Y. (2007). Evidentiary competence: sixth graders' understanding for gathering and interpreting evidence in scientific investigations. *Research in Science Education*, 37(1), 75–97.
- Jiménez-Aleixandre, M. P. (2007). Designing argumentation learning environments. In S. Erduran & M. P. Jiménez-Aleixandre (Eds.), *Argumentation in science education* (pp. 91–117). Dordrecht: Springer Netherlands.
- Jiménez-Aleixandre, M. P., & Erduran, S. (2007). Argumentation in science education: an overview. In M. P. Jiménez-Aleixandre & S. Erduran (Eds.), *Argumentation in science education. Perspectives from classroom-based research* (pp. 3–29). Dordrecht: Springer.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251.
- 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.
- Katchevich, D., Hofstein, A., & Mamlok-Naaman, R. (2013). Argumentation in the chemistry laboratory: inquiry and confirmatory experiments. *Research in Science Education*, 43(1), 317–345.
- Keller, J., Bohner, G., & Erb, H.-P. (2000). Intuitive und heuristische Urteilsbildung - verschiedene Prozesse? (Intuitive and heuristic decision-making—different processes?) *Zeitschrift für Sozialpsychologie*, 31(2), 87–101.
- Kelly, G. J., Druker, S., & Chen, C. (1998). Students' reasoning about electricity: combining performance assessments with argumentation analysis. *International Journal of Science Education*, 20(7), 849–871.
- Kelly, G. J., Crawford, T., & Green, J. (2001). Common task and uncommon knowledge: dissenting voices in the discursive construction of physics across small laboratory groups. *Linguistics and Education*, 12(2), 135–174.
- Kelly, G. J., Regev, J., & Prothero, W. (2007). Analysis of lines of reasoning in written argumentation. In S. Erduran & M. P. Jiménez-Aleixandre (Eds.), *Argumentation in science education. Perspectives from classroom-based research* (pp. 3–29). Dordrecht: Springer.
- Khine, M. S. (Ed.). (2012). *Perspectives on scientific argumentation*. Dordrecht: Springer Netherlands.
- Kind, P. M. (2013). Establishing assessment scales using a novel disciplinary rationale for scientific reasoning. *Journal of Research in Science Teaching*, 50(5), 530–560.
- Kind, P. M., Kind, V., Hofstein, A., & Wilson, J. (2011). Peer argumentation in the school science laboratory—exploring effects of task features. *International Journal of Science Education*, 33(18), 2527–2558.
- Klahr, D. (2000). *Exploring science: the cognition and development of discovery processes*. Cambridge: MIT Press.
- Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94(2), 211–228.
- Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). London; New York: Routledge.
- KMK: Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Ed.). (2004). *Bildungsstandards in Fach Physik für den Mittleren Schulabschluss. (German science standards for physics in middle schools)*. München: Wolters Kluwer.
- Kolstø, S. D., & Ratcliffe, M. (2007). Social aspects of argumentation. In M. P. Jiménez-Aleixandre & S. Erduran (Eds.), *Argumentation in science education. Perspectives from classroom-based research* (pp. 117–136). Dordrecht: Springer.
- Lee, H.-S., Liu, O. L., Pallant, A., Roohr, K. C., Pryputniewicz, S., & Buck, Z. E. (2014). Assessment of uncertainty-infused scientific argumentation. *Journal of Research in Science Teaching*, 51, 581–605.
- Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse (Test-development and test-analysis)*. Weinheim: Beltz, Psychologie Verl.-Union.
- Lin, J.-Y. (2007). Responses to anomalous data obtained from repeatable experiments in the laboratory. *Journal of Research in Science Teaching*, 44(3), 506–528.
- Lubben, F., Sadeck, M., Scholtz, Z., & Braund, M. (2010). Gauging students' untutored ability in argumentation about experimental data: a South African case study. *International Journal of Science Education*, 32(16), 2143–2166.

- Ludwig, T. (2017). Argumentieren beim Experimentieren - Die Bedeutung personaler und situationaler Faktoren (Dissertation) (Argumentation during experimentation—The influence of personal and situational factors). Humboldt-Universität zu Berlin.
- Ludwig, T., & Priemer, B. (2017). Nicht-erwartete Messdaten und experimentelle Beobachtungen im Physikunterricht - eine gute Gelegenheit im Unterricht zu argumentieren (Anomalous data and experimental observations in physics classes—a good opportunity to argue). *Praxis der Naturwissenschaften - Physik in der Schule*, *66*(3), 23–27.
- Ludwig, T., & Priemer, B. (2013). Argumentationen bei nicht-hypothesenkonformen Ergebnissen in Real- und Simulationsexperimenten (Argumentation from anomalous data from real and virtual experiments). In: S. Bernholt (Hrsg.), *Inquiry-based Learning - Forschendes Lernen* (Bd. 33, S. 731–733). Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Hannover 2012, Kiel: IPN.
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149.
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*(3), 519–530.
- Mason, L. (2001). Responses to anomalous data on controversial topics and theory change. *Learning and Instruction*, *11*(6), 453–483.
- McDonald, S. P., & Kelly, G. J. (2012). Beyond argumentation: sense-making discourse in the science classroom. In M. S. Khine (Ed.), *Perspectives on scientific argumentation* (pp. 265–281). Dordrecht: Springer.
- McNeill, K. L., & Berland, L. K. (2015). Design heuristics to enable students productive use of evidence in k-12 classrooms. Presented at the annual meeting of the National Association for Research in Science Teaching, Chicago, IL.
- McNeill, K. L., & Krajcik, J. (2007). Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations. In M. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 233–265). Taylor & Francis Group.
- Morse, J. M., & Niehaus, L. (2009). *Mixed method design: principles and procedures*. Walnut Creek: Left Coast Press.
- Nederhof, A. J. (1985). Methods of coping with social desirability bias: a review. *European Journal of Social Psychology*, *15*(3), 263–280.
- Nelson, R. A., & Olsson, M. G. (1986). The pendulum—rich physics from a simple system. *American Journal of Physics*, *54*(2), 112–121.
- NGSS Lead States. (2013). *Next generation science standards: for states, by states*. Washington: The National Academies Press.
- Njoo, M., & de Jong, T. (1993). Exploratory learning with a computer simulation for control theory: learning processes and instructional support. *Journal of Research in Science Teaching*, *30*(5), 821–844.
- Osborne, J., Erduran, S., Simon, S., & Monk, M. (2001). Enhancing the quality of argument in school science. *School Science Review*, *82*(301), 63–70.
- Osborne, J., Simon, S., Christodoulou, A., Howell-Richardson, C., & Richardson, K. (2013). Learning to argue: a study of four schools and their attempt to develop the use of argumentation as a common instructional practice and its impact on students. *Journal of Research in Science Teaching*, *50*(3), 315–347.
- Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123–205). Orlando: Academic Press.
- PhET Interactive Simulations. (2011). *Pendulum lab*. Boulder: University of Colorado Retrieved from <http://phet.colorado.edu/> Accessed 23 November 2012.
- Posner, G. J., Strike, K. A., Hewson, P., & Gertzog, W. A. (1982). Accommodation of a scientific conception: toward a theory of conceptual change. *Science Education*, *66*(2), 211–227.
- Priemer, B., & Hellwig, J. (2018). Learning about measurement uncertainties in secondary education: a model of the subject matter. *International Journal of Science and Mathematics Education*, *16*(1), 45–68.
- Pui-Wa, L., & Qiong, W. (2012). Estimation in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 164–180). New York: Guilford Press.
- R Core Team. (2014). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing Retrieved from <http://www.R-project.org/> Accessed 20 June 2015.
- Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373.
- Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology*, *84*(5), 754–775.
- Riemeier, T., von Aufschnaiter, C., Fleischhauer, J., & Rogge, C. (2012). Argumentationen von Schülern prozessbasiert analysieren: Ansatz, Vorgehen, Befunde und Implikationen (Process-based analysis of

- students' argumentations: approach, evidence and implications). *Zeitschrift Für Didaktik Der Naturwissenschaften*, 18, 181–200.
- Roberts, R., & Gott, R. (2010). Questioning the evidence for a claim in a socio-scientific issue: an aspect of scientific literacy. *Research in Science & Technological Education*, 28(3), 203–226.
- Robinson, W. R. (2000). A view of the science education research literature: scientific discovery learning with computer simulations. *Chemical Education Today*, 77(1), 17–18.
- Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36.
- Ryu, S., & Sandoval, W. A. (2012). Improvements to elementary children's epistemic understanding from sustained argumentation. *Science Education*, 96(3), 488–526.
- Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: current perspectives and recommendations for future directions. *Science Education*, 92(3), 447–472.
- Sandoval, W. A. (2003). Conceptual and epistemic aspects of students' scientific explanations. *Journal of the Learning Sciences*, 12(1), 5–51.
- Sandoval, W. A. (2012). Situating epistemological development. In *Volume 1: Full Papers* (pp. 347–354). Sydney.
- Sandoval, W. A., & Çam, A. (2011). Elementary children's judgments of the epistemic status of sources of justification. *Science Education*, 95(3), 383–408.
- Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction*, 23(1), 23–55.
- Sandoval, W. A., & Morrison, K. (2003). High school students' ideas about theories and theory change after a biological inquiry unit. *Journal of Research in Science Teaching*, 40(4), 369–392.
- Sandoval, W. A., Greene, J. A., & Bråten, I. (2016). Understanding and promoting thinking about knowledge: origins, issues, and future directions of research on epistemic cognition. *Review of Educational Research*, 40(1), 1, 40.
- Scheiner, S. M. (2004). Experiments, observations and other kinds of evidence. In M. L. Taper & S. R. Lele (Eds.), *The nature of scientific evidence: statistical, philosophical, and empirical considerations* (pp. 51–71). Chicago: University of Chicago Press.
- Schulz, L. E., Goodman, N. D., Tenenbaum, J. B., & Jenkins, A. C. (2008). Going beyond the evidence: abstract laws and preschoolers' responses to anomalous data. *Cognition*, 109(2), 211–223.
- Scott, P., Asoko, H., & Leach, J. (2007). Student conceptions and conceptual learning in science. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 31–56). London: Routledge.
- Shepardson, D. P. (1999). The role of anomalous data in restructuring fourth graders' frameworks for understanding electric circuits. *International Journal of Science Education*, 21(1), 77–94.
- Sinatra, G. M. (2005). The warming trend in conceptual change research: the legacy of Paul R. Pintrich. *Educational Psychologist*, 40(2), 107–115.
- Toulmin, S. E. (2003). *The uses of argument*. New York: Cambridge University Press.
- van Joolingen, W. R., & de Jong, T. (1997). An extended dual search space model of scientific discovery learning. *Instructional Science*, 25(5), 307–346.
- Walton, D. N. (1996). *Argumentation schemes for presumptive reasoning*. Mahwah: Lawrence Erlbaum.
- Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, 46(1), 71–95.
- Weiner, I. B., Schinka, J. A., & Velicer, W. F. (2012). *Handbook of psychology, research methods in psychology*. Wiley.
- West, S. G., Curran, P. J., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, 1(1), 16–29.
- West, S. G., Aaron, T. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). New York: Guilford Press.
- Wilson, C. D., Taylor, J. A., Kowalski, S. M., & Carlson, J. (2010). The relative effects and equity of inquiry-based and commonplace science teaching on students' knowledge, reasoning, and argumentation. *Journal of Research in Science Teaching*, 47(3), 276–301.
- Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching*, 39(1), 35–62.

Affiliations

Tobias Ludwig¹ · **Burkhard Priemer**² · **Doris Lewalter**³

¹ Institute of Physics and Technical Education, Karlsruhe University of Education, Bismarckstr. 10, 76133 Karlsruhe, Germany

² Department of Physics - Physics Education, Humboldt-Universität zu Berlin, Newtonstr. 15, 12489 Berlin, Germany

³ TUM School of Education, Technical University Munich, Arcisstraße 21, 80333 Munich, Germany