Editors' Suggestion

Students' conclusions from measurement data: The more decimal places, the better?

Karel Kok,* Burkhard Priemer, and Wiebke Musold
Humboldt-Universität zu Berlin, Physics Education, Department of Physics, Newtonstraße 15–12489, Berlin, Germany

Amy Masnick
Hofstra University, Hauser Hall, Psychology Department Hempstead, New York 11549, USA

(Received 17 October 2018; published 7 January 2019)

In this study with 153 middle school students, we investigate the influence of the number of decimal places from the reading of a measurement device on students' decisions to change or keep an initial hypothesis about falling objects. Participants were divided into three groups, introduced to two experiments—the time it takes a free falling object with a zero, and a nonzero initial horizontal velocity to fall a certain distance—and asked to state a hypothesis that compares the falling times of the two experiments. We asked the participants whether they wanted to change or keep their initial hypothesis after they were provided with data sets. Members of each group were given the same number of measurements but with a different number of decimal places. Results show that for an increase in the number of decimal places, the number of participants switching from a false to a correct hypothesis decreases, and at the same time the number of students switching from a correct to a false hypothesis increases. These results indicate that showing more exact data to students—given through different resolutions of the measurement device—may hinder students' ability to compare data sets and may lead them to incorrect conclusions. We argue that this is due to students' lack of knowledge about measurement uncertainties and the concept of variance.

DOI: 10.1103/PhysRevPhysEducRes.15.010103

I. INTRODUCTION

Judging the quality of data is a core competence that students should have [1], and being able to interpret data is a skill that is growing more important in our technological society [2]. Hence, data evaluation is included in science standards in different countries (e.g., USA: NGSS, UK: Department for Education, GER: Kultusministerkonferenz, NLD: SLO [3–6]). For example, students should be able to use data as evidence to justify a claim or hypothesis. To make these justifications, students should have some level of data literacy, which is fundamental to scientific argumentation [7]. Otherwise, students—as novices in a field—may base their justification on nonrational arguments [8] like intuition, which can lead to weak learning outcomes [9]. This is not to say that intuition is always irrational, as Weber shows for experts in mathematics when they are creating a proof [10].

For (scientific) claims to be justified or, more generally, to construct an empirically sound argument, experiments are conducted to gather empirical data as evidence. For a robust justification, the data have to be analyzed and interpreted [11]. The relevance and quality of these justifications thereby depend on the quality of the data. However, students often experience difficulties in judging the quality of the data. This can be illustrated with the example that, when faced with anomalous data students often tend to disregard the data as evidence, and fall back to their prior beliefs [11,12]. The review paper by Garfield and Ben-Zvi [13] shows that students' understanding of statistical concepts is often overestimated. Students are able to perform calculations, but lack a conceptual understanding. Because of this lack of deeper background knowledge in analyzing data, students may rely on their limited conceptions (e.g., knowing how to calculate a mean) or construct intuitive knowledge (e.g., abstract cognitive structures like "more measurements are better" comparable to p-prims suggested by di Sessa [14]). Thus, it seems likely that statistical data features like the number of measurements, mean values, or variance and students' competencies to work with these variables influence students' conclusions. Masnick and Morris [15] looked at some aspects of students' ability to compare data sets. Students were shown two data sets with different sample

*karel.kok@physik.hu-berlin.de

size, means, variability within and between data sets, and varied the sample size of the data sets. Results show that students recognize differences in sample size, differences in the mean, and variability between data sets. But variability within the data set (variance) had little effect on the students' comparison of the data sets. The competence in judging data sets is reported to increase with age. This study shows that students, starting as early as third grade, can interpret basic statistical quantities like the mean, but the more complex quantities like variance need more attention.

Further, measurement uncertainties—that can be the source of variance in data—are rarely discussed in schools even though they are inherent to all measurements and cannot be neglected when analyzing data [16]. Consequently, research shows students' low competences in understanding measurement uncertainties [13,16–18]. Lubben *et al.* [19] have categorized students' reasoning with data into two categories: point paradigm and set paradigm. In the point paradigm, students regard each measurement as an isolated event, where repeated measures, when done right, result in the same true value. In the set paradigm, students see a series of measurements with random variation as a way to approach the true value. Allie *et al.* [20] report that most high school students are firmly located in the point paradigm, and that set-paradigm actions (such as calculating a mean) are mostly rote responses.

Given the fact that students have difficulties in dealing with uncertainties and at the same time construct intuitive knowledge (e.g., the more exact the data, the better the quality), how would they interpret data when both aspects are in conflict? The question we ask is as follows: what is the influence of the number of decimal places of measurements in a data set on students' decisions to keep or change their hypothesis when comparing data sets in the context of a physics problem? A typical idea a student may follow is that increasing the number of decimal places increases the exactness and, hence, the quality of a measurement ("the more the better"). However, more decimal places also make measurement uncertainties more obvious since the measurement results—the single numerical measurements that the students see—in the data set differ. Here, the increasing exactness leads to an apparent increase in variability, which may be confusing to students ("the more the worse"). So, does more exactness lead to more students choosing the correct hypothesis? Can students use the given exactness appropriately to draw the right conclusions?

With this study, we aim to answer the following question: What is the influence of the number of decimal places in the result of a physics experiment on school students' decision when reevaluating their initial hypothesis? And if the number of decimal places makes a difference in students' reevaluation of their hypotheses, do the methods of analyzing the data or the differences seen in data sets explain the effect? Finally, can the type of justifications students refer to predict their choice of the correct hypothesis?

II. METHOD

In this study, 153 participants of grades 8–10 (average age 14 years) from an urban high school in Germany took part (convenience sample). The school was chosen such that the participants were unfamiliar with the context of the study. All participants had good reading and writing skills. The experiment was done during the normal 45-min physics class, the participants had calculators available, and we have found no evidence of participants not being able to finish the questionnaire in time.

A. Introduction to the experiment

Prior to the questionnaire, the participants were shown a 4-min video showing and explaining the experimental setup, and how the measurement data were collected. One ball is attached to an electromagnet and is dropped when a switch is pressed. As the ball falls through a photogate, a computer starts a timer, the timer is stopped once the ball reaches the photogate at the bottom. The experiment is then repeated, but for a ball that is launched by letting it roll down an incline. At the end of the incline, at the same height as the dropped ball, a photogate is placed. Again, the computer measures the time it takes the launched ball to reach the second photogate at the bottom.

After seeing the video, the participants each got a questionnaire, in which the experimental setup is depicted (see Fig. 1) and the experimental procedure is repeated in writing. With the questionnaire, we randomly assigned the participants into one of three groups: A, B, and C. The participants are asked to form a hypothesis (multiple choice) as to which object has the longest falling time: the free falling object, the launched object, or that this is the same for both objects.

B. Showing the data set

After stating their initial hypothesis, the participants were shown results of the experiment. Depending on the group to which the participant was assigned (A, B, or C), the participant saw a data set with a different number of

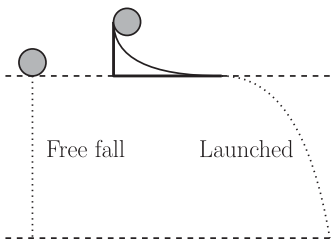

FIG. 1. A sketch of the experimental setup as shown to the participants.

TABLE I. The data sets of the two experimental settings as shown to the different groups A, B, and C. Times are in seconds.

Free fall			Launched		
A	B	C	A	B	C
0.53	0.535	0.5351	0.53	0.535	0.5350
0.53	0.535	0.5353	0.53	0.535	0.5352
0.53	0.535	0.5355	0.53	0.534	0.5347
0.53	0.534	0.5347	0.53	0.535	0.5354
0.54	0.534	0.5349	0.53	0.535	0.5351
0.53	0.534	0.5348	0.53	0.535	0.5352

decimal places, see Table I. Then, the participants had time to analyze their data.

After the data analysis, the participants were asked what strategy they would use to compare the two data sets. The eight multiple-choice options given were 1 = "compare the sums of the rows", 2 = "compare the means of the rows", 3 = "compare the medians", 4 = "compare the value that occurs most often (modus)", 5 = "calculate pairwise differences and compare all these differences", 6 = "compare the data sets step by step in a pairwise manner", 7 = "look at the rows and see if there is a difference", 8 = "none of the above". Options 1–5 indicate a response where the student has a clear, purposeful strategy in mind to compare the data. In contrast, options 6 and 7 are not as structured and purposeful. These answer options were based on responses during a pilot study, and resemble authentic student responses.

The participants were also asked (multiple choice) whether they saw a "clear difference," a "small difference," or "no difference" between the two data sets.

Then the participants were asked to reevaluate their hypothesis, again in multiple-choice form.

C. Reasoning and justification

We asked the participants to write down their reasoning for changing or not changing their hypothesis. This reasoning was then classified into different categories, as shown in the flowchart in Fig. 2. Quantitative reasons that entail explicit numerical mentions of values referring to specific data points, calculated means, total time, differences, etc. were classified as numerical (code NM). Reasons that mention words like: "clear," "few," "some," or other nonnumerical descriptions of the data were classified as non-numerical (code NN). We also distinguish a group of participants that did not mention the data but reasoned on a theoretical basis (code T), and we have identified a group of participants that neither reasoned in a numerical, non-numerical, nor in a theoretical manner (code 0). Lastly, there is a group that did not write down anything at all (code NA). Two raters used a coding manual and coded a sample of 20 responses ($\kappa = 0.93$), which gave only one discrepancy that was resolved in discussion. After that, one rater coded the remainder of the responses.

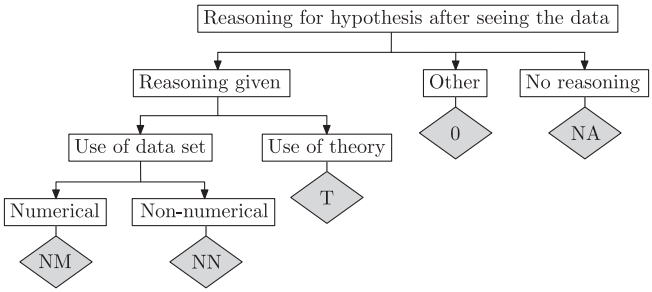

FIG. 2. The flow chart of classifications for the reasoning given by the participants when reevaluating their hypothesis after having seen the data.

Finally, we presented the participants with twenty statements about factors that might have had an influence on the reevaluation of their hypothesis. The participants were asked to which extent the statement applied in their decision to change or keep their hypothesis using a five-point Likert scale. The steps on the scale ranged from 1 = "(statement) does not apply", to 5 = "(statement) fully applies". These twenty items probe the influence of the following four factors on participants' justification in a self-assessment: data as evidence, measurement uncertainties, expert knowledge, and intuition. These items are thoroughly described by Ludwig [9].

III. RESULTS

Prior to showing the participants the data, 58% thought the launched ball would have the longest drop time, 33% of the participants thought the objects would land at the same time (correct solution), and 9% thought that the free-falling object would have the longest drop time. There were no significant differences between the groups A, B, and C at this time, $\chi^2(4) = 1.85$, $p > 0.1$. The distribution is similar to one of the groups described in Whitaker [21]. This satisfies our assumption of homogeneity among the three groups.

A. Reevaluation of the hypothesis

After showing the data sets to the different groups, we asked the participants to reevaluate their hypothesis. We made five different categories of how participants change their hypothesis; correct to correct (CC), correct to false (CF), false to correct (FC), false to other false (FoF), and false to the same false (FF). The percentage distribution of the participants in the three groups (A, B, and C) is shown in Table II. We see that the distribution of the number of participants over these different categories differs significantly between groups and has a medium effect strength, $\chi^2(8) = 15.55$, $p < 0.05$, $w = 0.32$.

In group A (two decimal places), 40% of the participants switched from a false to the correct hypothesis; in groups B (three decimal places) and C (four decimal places) this percentage is lower at 31% and 33%, respectively. On the

TABLE II. Percentage of participants switching their hypothesis after having seen the data. The different categories are labeled: CC = kept the correct hypothesis, CF = switched from the correct to a false hypothesis, FC = switched from a false to the correct hypothesis, FoF = switched from a false to the other false hypothesis, FF = kept the same false hypothesis.

	Group		
Category	A ($n = 52$)	B ($n = 52$)	C ($n = 49$)
CC	35	25	18
CF	0	9	10
FC	40	31	33
FoF	13	27	16
FF	12	8	23

other hand, no participants in group A switched from a correct to a false hypothesis, whereas in both group B and C, about 10% of the participants switched from a correct to a false hypothesis.

We thus conclude that when we increase the number of decimal places of a given data set, the number of students switching from a false to a correct hypothesis decreases, and at the same time triggers some students to switch from a correct to a false hypothesis.

B. Participants' data analyzing methods and perceived differences in the data sets

Can this result be explained by different data analysis methods in the three groups? Or do students' views regarding the differences they see in the data sets vary between the groups? When asked about the strategy to compare the data, we found no significant differences between the groups for the distribution of these strategies, $\chi^2(12) = 16.58$, $p > 0.1$. An average of 76% of the participants chose to compare the sums of the rows or the means of the rows (options 1 and 2), none of the participants chose to compare the median (option 3), and 3%–7% chose the other options.

In participants' responses to the question whether they saw a clear, small or no difference between the data sets, we did not find any significant difference between groups, $\chi^2(4) = 4.62$, $p > 0.1$. In general, 83% of all participants report seeing a small difference between the data sets. This might be due to the fact that we made the answer option multiple choice, and that the participants each have different interpretations of the answer options clear and small.

We conclude that neither the data analysis method nor the perceived difference between the two data sets explains why the three groups differ in their choices of the post hypotheses.

C. Participants' quantitative and qualitative reasoning

For the classification of participants' reasoning for the reevaluation of their hypothesis, we looked at the difference in quantitative reasoning, and qualitative reasoning. We consider all the numerical reasons (code NM) as quantitative, and all the non-numerical, and theory-based reasons (code NN and T) as qualitative. Table III shows the percentage distribution over these two categories for groups A, B, and C. Four participants had other reasons (code 0), and five participants did not give a reason for their hypothesis (code NA). These participants were taken out of the statistics. In total there were 32 participants that wrote down a quantitative reason, and 112 participants that wrote down a qualitative reason.

TABLE III. Percentage of participants in the different groups that base the reasoning of their hypothesis on quantitative (code NM) or qualitative (code NN and T) grounds.

	A ($n = 50$)	B ($n = 48$)	C ($n = 46$)
Quantitatively ($n = 32$)	32	17	17
Qualitatively ($n = 112$)	68	83	83

We see that the distribution of responses in groups B and C are identical, and the number of participants that base their reasoning on a quantitative explanation is more than half of that in group A. When the results of groups B and C are added and compared with group A, we see a significant difference between groups A and B + C, $\chi^2(1) = 4.24$, $p < 0.05$.

Of the participants that wrote down a quantitative reason, 84% end up making the right choice for the reevaluation of their hypothesis. For those who gave a qualitative reason, this is only 54%; see Table IV. The difference between the quantitative and qualitative group is significant, $\chi^2(2) = 10.13$, $p < 0.01$.

From this, we conclude that increasing the number of decimal places shifts the students' thinking to a more qualitative way of data perception and comparison. This shift, in turn, leads to a worse judgment in data comparison.

D. Influencing factors on the justification of the hypothesis

Lastly, we looked at four factors (data as evidence, measurement uncertainties, intuition, and expert knowledge), and to what degree they have influenced the participant's decision to reevaluate their hypothesis. We find a negative correlation of medium strength between data as evidence, and intuition-based justification in group B, $r = -0.40$, $p < 0.01$. For group C we find the

TABLE IV. Percentage distribution of participants' reevaluated hypotheses for quantitative (code NM) and qualitative (code NN and T) reasoning.

	Quantitatively ($n = 32$)	Qualitatively ($n = 112$)
Same	84	54
Free fall	6	30
Launched	10	16

TABLE V. The logistic regression model for changing to a correct hypothesis for the predictors data as evidence and measurement uncertainties. The odds show that the chance of changing from a false to the correct hypothesis increase when the justification is based on these predicting factors. $*p < 0.05$.

			Odds ratio		
Factor	B	SE	Lower	Center	Upper
Data as evidence	0.59*	0.25	1.11	1.80	3.03
Measurement uncertainty	0.51*	0.22	1.10	1.66	2.59

same negative correlation but with a large strength, $r = -0.54$, $p < 0.001$. This correlation does not emerge in group A, $r = -0.09$, $p > 0.1$. This means that, when more exact data are shown to students, the ones that rely on data as evidence for their justification do not rely on their intuition, and vice versa.

For participants who had a false initial hypothesis, we looked at which factors, influencing the justification, had a positive influence on the chances of participants changing to the correct hypothesis. To do this we conducted a logistic regression model and excluded non-significant factors in a step-by-step manner. We found that a higher score on the factors data as evidence and measurement uncertainties, leads to an increase in the odds of changing to the right hypothesis; see Table V. For the factor data as evidence, this means that a one-step increase on the Likert scale increases the odds of changing to a correct hypothesis by 1.80 times. Both effects are small but significant, $p < 0.05$.

IV. DISCUSSION

The results of our study indicate that the number of decimal places in quantitative experimental data reduces students' ability to critically compare data sets. This, in turn, may hinder students' learning of physics content. Better measurement equipment—that leads to more exactness by means of more decimal places in the measurement results and hence stronger evidence—can lead some students to reject a correct hypothesis.

How can this be explained? While 76% of the participants indicated they compared the means or the sums of the rows of their two data sets, they knew no method to judge if the calculated difference between the two is relevant. Hence, one possibility is that they disregarded the data and made their decision based on theoretical or everyday life beliefs. Another explanation is that they stuck to the quantitative data and made their decision whether to change or keep their initial hypothesis solely by comparing two numbers. If the numbers differed, participants assumed that there is a main effect.

This is in line with a study conducted by Priemer and Hellwig [16], in which students measured temperatures inside foam cubes of different sizes and used only the readings of the thermometer with all its digits to investigate if there are differences between the temperatures. One of the students' lines of argument was that if the numbers differ, there must be a main effect. Thus, they came to the wrong conclusion that the size of the foam cubes influences the temperature inside. This is in line with our finding that the majority of students see a small difference between the data sets. Of course the terms clear, and small are subjective terms, but the students' responses are independent of the number of decimal places. Also, the reevaluation of their hypothesis indicates that the students have no way of determining whether this small difference is relevant.

So, how do the students reason? We see that, when we increase the number of decimal places, a negative correlation between the factors "data as evidence" and "intuition"—factors that students reported as having influenced their justification—emerges: The less students argue with evidence, the more they base their decision on intuition. So, if we use our result that more exact data lead to more students who put empirical evidence aside and use qualitative justifications, we can conclude that this leads to more intuitive decisions. This pattern in turn supports our claim that showing more exact data to students reduces their ability to critically compare data sets. This is also supported by our finding that students who reason in a quantitative manner are significantly better at the reevaluation of their hypothesis than the students that answer in a vague, qualitative manner.

Furthermore, students that are able to justify the reevaluation of their hypothesis based on data as evidence (and measurement uncertainties), have increased odds of switching to the correct hypothesis. We believe that these students are better at comparing the data and can incorporate this in their justification without having to rely on their intuition.

Why is it that students who know how to calculate means appear to be so overwhelmed by the data that they cannot gauge the significance of the difference between the means and fall back on their intuition? We think that the absence of the concept of (some measure of) variance hinders the students to put the difference into perspective. Students may not be surprised by varying measurement results under changing conditions, e.g., temperatures during a year. However, in our setting they might not even expect a spread in measurement results, because the repeated measurements were taken under identical conditions. We already saw that students who are able to justify the reevaluation of their hypothesis based on measurement uncertainties, have increased odds of switching to the correct hypothesis. These students rely on the measurement uncertainties for their justification, making it an integral part of their argumentation. They might be able to compare the two mean values, recognize that their difference falls within the range of the measurement uncertainties, and conclude that there exists no (significant) difference between the data sets. With that, they have incorporated variance into their argumentation, complementary to the mean value.

V. IMPLICATIONS FOR TEACHERS

The result that more exact data decrease students' ability to judge the quality of data, and consequently the evidence for (scientific) claims, is somewhat unsettling. One obvious and seemingly "easy" way of helping students in tasks that involve the comparison of data sets would be to simply reduce the number of decimals. Our study shows that this will increase students' achievement. This, however, ignores the fact that students will encounter data of any exactness in their everyday lives. But, more important, this only hides measurement uncertainties away from the students by making uncertainties "invisible." Clearly, the uncertainties still exist. Using the convention that the last digit of a digital measurement device indicates the uncertainty of this device (no matter how many measurements are made), we can state that decreasing the number of decimal places on the device increases the measurement uncertainty. This is because the statistical variance (in the form of random fluctuations in the measurements) becomes smaller than the uncertainty of the device (for two decimal places this is 0.01 s, for three decimal places 0.001 s, etc.). So, giving students less exact data not only increases the measurement uncertainty but also misleads students to believe that the uncertainties disappear. Our participants in group A—those who worked only with two decimal places—may have chosen the right answer for a wrong reason. Assuming an absence of uncertainties (or not being aware of uncertainties at all), they judged their data as "perfect" and based their decision on this condition. In contrast, the students that realized a variance in the data—statistical fluctuations—became hesitant and more critical concerning the quality of the data. For this reason, some of the students choose not to use the quantitative data at all but made their decision based on qualitative justifications. Because of very limited knowledge here, they had to rely on intuition or vague everyday life concepts that lead to wrong conclusions. Thus, we must teach the students the necessary skills to compare data to enable them to make use of more precise measurements and high quality data.

To improve this skill we suggest that students will have to supplement their notion of the result—which now only consists of the mean value—with some form of a confidence interval. Without this interval around the mean value, students will have no way of judging the (in) significance of the difference between the mean values, other than their intuitive response. Since students often experience difficulties when thinking about measurement uncertainties, e.g., [13,22], care has to be taken in designing suitable tasks.

In order to do this, we suggest that teachers should design tasks that increase students' awareness and understanding of measurement uncertainties and variance. First students should be made aware of the unavoidable presence of measurement uncertainties. This would, in turn, create a need for some confidence interval around the mean, which can be quantified. This interval can be a rudimentary quantity like minimum-maximum values, percentage spread around the mean, or the spread between the middle n data points. More advanced students could even go as far as to calculate the more formal variance or standard deviation. The next step is to make this interval an integral part of the result, complementary to the mean. With that, students can compare two measurement results by looking at the degree of overlap between the confidence intervals. We believe that this understanding of variance and measurement uncertainties will allow students to reason based on the data in a quantitative manner, and lead them away from intuitive responses.

There are several examples in science contexts to be found, e.g., Refs. [16,23,24]. But teaching about measurement uncertainties and variance is not a task reserved just for science teachers. Since data, and judging the quality of this data, is becoming so prominent in our everyday lives, teachers in all subjects should try to incorporate this into their classes. This can be done in a quantitative way of calculation, but also in a qualitative way by raising awareness and discussing the limits of results, and the validity of claims.

ACKNOWLEDGMENTS

We acknowledge support by the German Research Foundation (DFG) and the Open Access Publication Fund of Humboldt-Universität zu Berlin.

[1] C. A. Chinn and B. A. Malhotra, Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks, Sci. Educ. **86**, 175 (2002).

[2] S. V. Sharma, High school students interpreting tables and graphs: Implications for research, Int. J. Sci. Educ. **4**, 241 (2006).

[3] NGSS Lead States, *Next Generation Science Standards: For States, By States* (The National Academies Press, Washington, DC, 2013).

[4] Department for Education, Science Programmes of Study: Key Stage 4, Report No. DFE-00677-2014, 2014.

[5] Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, Bildungsstandards Im Fach Physik Für Den Mittleren Schulabschluss, Wolters Kluwer, Report No. 06218, 2004.

[6] SLO–Nationaal Expertisecentrum Leerplanontwikkeling, Kennisbasis Natuurwetenschappen En Technologie Voor de Onderbouw vo: Een Richtinggevend Leerplankader,

Nationaal Expertisecentrum Leerplanontwikkeling, Report No. 4.6691.552, 2014.

[7] K. L. McNeill and J. Krajcik, Middle School Students' Use of Appropriate and Inappropriate Evidence in Writing Scientific Explanations, in *Thinking with Data*, edited by M. Lovett and P. Shah (Taylor & Francis Group, LLC, New York, 2007), pp. 233–267.

[8] J. A. Dole and G. M. Sinatra, Reconceptalizing change in the cognitive construction of knowledge, Educ. Psych. **33**, 109 (1998).

[9] T. Ludwig, Argumentieren Beim Experimentieren in Der Physik, Ph.D. thesis, Humboldt University, Berlin, 2017.

[10] K. Weber, Beyond Proving and Explaining: Proofs That Justify the Use of Definitions and Axiomatic Structures and Proofs That Illustrate Technique, For the Learning of Mathematics **22**, 14 (2002).

[11] D. Klahr and K. Dunbar, Dual space search during scientific reasoning, Cogn. Sci. **12**, 1 (1988).

[12] Z. Kanari and R. Millar, Reasoning from data: How students collect and interpret data in science investigations, J. Res. Sci. Teach. **41**, 748 (2004).

[13] J. Garfield and D. Ben-Zvi, How students learn statistics revisited: A current review of research on teaching and learning statistics, Int. Stat. Rev. **75**, 372 (2007).

[14] A. A. diSessa, in *Mental Models*, edited by D. Gentner and A. L. Stevens (Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1983), pp. 15–34.

[15] A. M. Masnick and B. J. Morris, Investigating the development of data evaluation: The role of data characteristics, Child Development **79**, 1032 (2008).

[16] B. Priemer and J. Hellwig, Learning about measurement uncertainties in secondary education: A model of the subject matter, Int. J. Sci. Math. Educ. **16**, 45 (2016).

[17] A. Buffler, S. Allie, and F. Lubben, The development of first year physics students' ideas about measurement in terms of point and set paradigms, Int. J. Sci. Educ. **23**, 1137 (2001).

[18] F. Lubben and R. Millar, Children's ideas about the reliability of experimental data, Int. J. Sci. Educ. **18**, 955 (1996).

[19] F. Lubben, B. Campbell, A. Buffler, and S. Allie, Point and set reasoning in practical science measurement by entering university freshmen, Sci. Educ. **85**, 311 (2001).

[20] S. Allie, A. Buffler, F. Lubben, and B. Campbell, in *Research in Science Education—Past, Present, and Future*, edited by H. Behrendt, H. Dahncke, R. Duit, W. Gräber, M. Komorek, A. Kross, and P. Reiska (Kluwer Academic Publishers, Dordrecht, 2002), pp. 331–336.

[21] R. J. Whitaker, Aristotle is not dead: Student understanding of trajectory motion, Am. J. Phys. **51**, 352 (1983).

[22] A. M. Masnick and D. Klahr, Error matters: An initial exploration of elementary school children's understanding of experimental error, J. Cognit. Dev. **4**, 67 (2003).

[23] A. Buffler, S. Allie, and F. Lubben, Teaching measurement and uncertainty the GUM way, Phys. Teach. **46**, 539 (2008).

[24] A. J. Petrosino, R. Lehrer, and L. Schauble, Structuring error and experimental variation as distribution in the fourth grade, Math. Think. Learn. **5**, 131 (2003).