DECISION-MAKING IN UNCERTAINTY-INFUSED LEARNING SITUATIONS WITH EXPERIMENTS IN PHYSICS CLASSES

Tobias Ludwig[1], Burkhard Priemer[1], and Doris Lewalter[2]
[1]Humboldt-Universität zu Berlin, Institut für Physik, Newtonstr. 15, 12489 Berlin, Germany
[2]TUM School of Education, Arcisstraße 21, 80333 München, Germany
Tobias.Ludwig@physik.hu-berlin.de

Evaluating experimental data calls for dealing with measurement uncertainty (MU); this is often reported as a learning obstacle to scientific concepts in lab work. We report on a study of 1,545 secondary school students who evaluated self-collected data in a physics lab. Results reveal a) that students are aware of MUs depending on data variability and b) that the odds of framing a correct hypothesis after the experiment decrease (OR=0.66[0.57;0.76], p<.001) when students are aware of MUs in their decision-making process. This contradicts studies that identified students' lack of awareness of MUs as an obstacle in the evaluation of experimental data, and it could be interpreted to mean that learners decide conservatively when evaluating uncertainty-infused learning situations because of a lack of knowledge about MUs.

INTRODUCTION

The evaluation of data and the justification of scientific hypotheses are at the core of 'doing science' and accordingly implemented in many school science standards (e.g., NGSS Lead States, 2013). However, the evaluation of quantitative data often calls for dealing with measurement uncertainties, a fact often neglected in schools (Priemer & Hellwig, 2016). While previous research has shown that students are able generally to comprehend concepts of uncertainties and discuss sources of uncertainty (Masnick & Klahr, 2003), students need a certain degree of unambiguity in the data for the revision of prior beliefs (Kanari & Millar, 2004), as they fail otherwise to adequately reason from data while generating a claim from data. In the context of lab work, it is particularly relevant to investigate how students justify a scientific claim based on experimentally gained data, which are often infused with uncertainty. Kanari and Millar (2004, p. 767) suggest that, besides the ability to deal with it appropriately, "the difficulty lies in students' awareness that all measurements are inevitably subject to uncertainty." In this context, little is known about how students' awareness of measurement uncertainties influences reasoning from data. We want to contribute to filling this gap by investigating how awareness of data uncertainties influences students' decision-making in hypothesis testing in a physics context (RQ 1). Furthermore, previous research shows that students' reasoning from data might be influenced by data quality, e.g., the number of significant digits or variability (Masnick & Morris, 2008). For this reason, we also investigate the influence of data quality on students' awareness of measurement uncertainties (RQ 2).

RESEARCH QUESTIONS

1. How does the degree to which students show awareness of data uncertainty influence the choice of a scientific hypothesis in terms of its correctness?
2. How does the quality of experimentally gained data (herein, the number of significant places, more or less data variability) influence students' awareness of measurement uncertainties and their use of data as evidence in hypothesis testing?

METHODS

The study was conducted among 1,545 secondary school students aged 13–16 years. Students were given a practical task to frame a hypothesis about the relation between the mass of a pendulum and its period of oscillation and to test this hypothesis through experimentation. The examination of the relation between pendulum mass and swing time is an easy experimental task, and it provides the opportunity for quantitative data collection. This context allows the participants to generate a subjectively plausible hypothesis formed from everyday life experiences. In the context of the pendulum, one can observe a variety of common misconceptions (Kanari & Millar, 2004). Accordingly, 87.4% of participants framed an incorrect hypothesis and will be confronted with anomalous data when conducting the experiment. We use this conflicting situation to engage students

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org [© 2018 ISI/IASE]

in a profound data evaluation process in which students must decide whether to keep their (incorrect) initial hypothesis (e.g., pendulum mass affects swing time) or change it in favor of a correct one (time of swing is unaffected by pendulum mass). We examined the justifications for this decision using a well-elaborated instrument consisting of Likert-scaled items. The steps on the Likert scale ranged from 1 = "(statement) does not apply" to 5 = "applies fully." In this test, which is answered immediately following experimentation, participants are asked to state to what extent the statements apply in their argumentation supporting or rejecting their hypotheses (for details regarding the questionnaire see Ludwig, 2017; Ludwig & Priemer, 2014). While this questionnaire consists of four different subscales, only two are included in the following analysis. One subscale included in this analysis is "Use of Data as Evidence," which probes to what extent students rely on the measurement data as evidence when making decisions regarding their incorrect initial hypotheses (example item: "The measurement results were the biggest factor in my decision"). Furthermore, the awareness of measurement uncertainties was assessed by the scale "Measurement Uncertainties (explicit)" (example item: "When making my decision, I took into account that the experiment is not exact"). Concerning research question 2, participants were assigned randomly to three different conditions, which allowed students to generate data of different quality.

Students in group A worked with hands-on material to test their hypothesis. The pendulum length was about 0.8–1 m, and the stopwatch provided allowed the swing time to be measured to one-hundredths of a second. The biggest source of uncertainty in this group is human reaction time, which can be assumed to be 0.5 s. Students in group B worked with a computer-simulated environment. Again, a digital stopwatch, which had to be started manually, was provided, allowing for swing time to be measured to one-hundredths of a second. Different from group A, the experimental demand is easier in group B. For example, the pendulum can be stopped and started by clicking a button. Furthermore, we decided to increase the length of the pendulum to 2–3 m, allowing students to observe the phenomenon on a standard computer monitor because the angular velocity is slower. This, in turn, increases the swing time and lowers the relative measurement uncertainty (because the reaction time stays the same). Group C worked with the same computer simulation as group B, the only difference being the stopwatch, which works automatically by means of an optical barrier, presents the measurement with four significant digits. Because the stopwatch is measuring automatically, the precision of the measurement in group C is indefinitely high. As well, because of the characteristics of the experiments outlined above, the uncertainty in the measurement data decreased from group A to C, while the data quality increased. The students' hypotheses before and after the experiment were coded as binary (1 = correct).

Data analyses were carried out by means of structural equation modeling using the package lavaan in R (Rosseel, 2012) and Mplus (Muthén & Muthén, 1998). All participants with a correct initial hypothesis were excluded. In total, 1,351 students entered the analysis. The overall measurement model, consisting of two latent variables representing the underlying two scales, showed an excellent fit to the data ($\chi^2(33) = 153.34$, $p < .01$, CFI = .96, RMSEA = .052, SRMR = .05). The model-based reliability for both variables are .75 for the measurement uncertainties scale and .85 for the data as evidence scale.

RESULTS

A descriptive analysis of the manifest scale means revealed apparently no difference in the use of data as evidence among groups A, B and C in the students' justification for or against a hypothesis (Table 1). However, the means of the measurement uncertainties scale tended to decrease when data variability decreased (from group A to C). The statistical significance of between-group differences was tested within a multi-group structural equation model (Kline, 2016). Because the questionnaire is inextricably connected to the process of experimentation across all groups, special focus was placed on the establishment of measurement invariance (MI) as a prerequisite necessary for comparing means across groups (Vandenberg & Lance, 2000). MI could be established to the level of scalar MI for both sub-scales. The latent means of both scales were fixed to zero in group A, while the other means were freely estimated. Again, this model showed a very good fit to the data ($\chi^2(131) = 307.4$, $p < .01$, CFI = .95, RMSEA = .055, SRMR = .06). Standardized mean differences (interpretable as Cohen's *d)* are also shown in Table 1. The latent means differ significantly across groups A–C for the measurement uncertainties scale, while there is no substantial difference for the data as evidence scale. The effect size between groups A and B is small, while the difference between groups A and C and between groups B and C can be considered large. As the correctness of the hypothesis after the experiment is coded in binary, a multiple logistic regression analysis was carried out to regress the outcome of the chosen hypothesis after the experiment on the measurement uncertainties and data as evidence scales. The model reveals that both variables influence significantly the correctness of the hypothesis after the experiment. The odds ratios for the data as evidence scale is 3.40 [2.79;4.12], which means that the odds for framing a correct hypothesis for every 1-point increase on the original scale are multiplied by 3.40. This can be seen as a medium-to-large effect. The odds ratio for the scale measurement uncertainties decreases the odds for framing a correct hypothesis by 0.66 [0.57;0.76], which can be interpreted as a small effect (Olivier, May, & Bell, 2016).

Table 1: Proportion of correct hypotheses after the experiment, scale means, and standardized mean differences for the measurement uncertainties and data as evidence scales aggregated by the experimental condition.

	Group A, hands-on, manual stopwatch, n = 697	Group B, computer-simulated experiment, manual stopwatch, n = 260	Group C, computer-simulated experiment, automatic stopwatch, n = 394
Incorrect hypothesis before the experiment	100 %	100 %	100 %
Correct hypothesis after the experiment	34.5 %	49.6 %	74.4 %
Scale means in Likert-scale units (SD)			
Data as Evidence	3.97 (0.76)	3.98 (0.89)	3.99 (0.89)
Measurement Uncertainties	3.43 (0.72)	3.27 (0.81)	2.69 (0.83)
Standardized mean differences			
Data as Evidence	0	0.02	0.09
Measurement Uncertainties	0 (reference group)	-0.25*	-1.13*

*mean difference significantly differed from the reference group, p < .05

CONCLUSION

The main goal of this paper was to investigate the interplay of students' awareness of measurement uncertainties, data quality, and hypothesis testing in the physics lab. On the one hand, our results imply that students are able to identify measurement uncertainties as a relevant issue in reasoning from data, as students show awareness of measurement uncertainties by performing relatively high on the 5-step Likert scale (between 2.69 and 3.43, depending on the group). This result seems to contradict that of Kanari and Millar (2004), who reported that students lack awareness of measurement uncertainties while analyzing data might hinder them in making correct inferences about variables in the pendulum context. However, equivalent to Kanari and Millar's work, our students also had difficulties framing a correct hypothesis after the experiment. To identify possible causes for why students tend to maintain a wrong hypothesis after the experiment, we carried out a logistic regression analysis in which we predicted the outcome of the hypothesis after the experiment by means of the scales "Awareness of Measurement Uncertainties" and "Use of Data as Evidence." While there are of course more predictors for framing a correct hypothesis, awareness of measurement uncertainties while justifying a hypothesis has a significant medium-sized negative effect on the correctness of the hypothesis after the experiment. We interpret this result so that students behave *conservatively* in making statistical decisions in uncertainty-infused situations: while they obviously perceive data as uncertain, they maintain their initial wrong hypothesis, because they seem to lack knowledge of how to deal appropriately with uncertainties in their measurements. In contrast to Kanari and Millar's work, this is not because they are *unaware* of uncertainties. This is strong evidence for the claim that we do need to foster explicitly students' abilities in evaluating data and uncertainties in school science labs.

With regard to research question 2, our results show that with an increase in data quality (less variability and more significant places), students perform lower on the measurement uncertainties awareness scale and perform better in framing a correct hypothesis after the experiment. This can be interpreted in two ways. First, this result implies we need to be aware of the fact that different kinds of measurement devices, providing data of different quality, may lead to different learning outcomes, because students may or may not be able to analyze data appropriately. Second, these results support similar research that showed learners' data evaluation can be influenced by certain characteristics (Masnick & Morris, 2008). Although all three learning environments in this study allowed learners to come to the same conclusion (no relation between swing time of swing and pendulum mass), participants performed significantly lower when data seemed to show greater variability. Similar to Masnick and Morris' (2008) suggestion, it seems that students tend to evaluate data means and variances intuitively rather than explicitly. As we lowered the variability from group A to C, we could observe an increase in correct hypotheses after the experiment. This may lead to the conclusion that an intuitive evaluation is only successful if data variability is minimal.

ACKNOWLEDGEMENT

This research was partially funded by the Deutsche Telekom Stiftung and partially funded by the German Research Foundation (DFG).

REFERENCES

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.

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling* (4th ed.). The Guilford Press.

Ludwig, T. (2017). Argumentieren beim Experimentieren - Die Bedeutung personaler und situationaler Faktoren (Dissertation). Humboldt-Universität zu Berlin, Berlin.

Ludwig, T., & Priemer, B. (2014). Ein Instrument zur Erfassung von Argumentationen beim Experimentieren. In S. Bernholt (Ed.), Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht (Vol. 34, pp. 273–275). Kiel: IPN.

Masnick, A. M., & Klahr, D. (2003). Error Matters: An Initial Exploration of Elementary School Children's Understanding of Experimental Error. Journal of Cognition and Development, *4*(1), 67–98.

Masnick, A. M., & Morris, B. J. (2008). Investigating the Development of Data Evaluation: The Role of Data Characteristics. *Child Development*, 79(4), 1032–1048.

Muthén, B. O., & Muthén, L. K. (1998). *Mplus User's Guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. Retrieved from nextgenscience.org

Olivier, J., May, W. L., & Bell, M. L. (2016). Relative effect sizes for measures of risk. *Communications in Statistics - Theory and Methods*, 46(14), 6774-6781.

Priemer, B., & Hellwig, J. (2016). Learning About Measurement Uncertainties in Secondary Education: A Model of the Subject Matter. *International Journal of Science and Mathematics Education*, 16(1), 45-68.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36.