Karel Kok[1]
Burkhard Priemer[1]

[1]Humboldt-Universität zu Berlin

Evaluating students' justifications in a data comparison problem

State of the Field

Judging the quality of (experimental) data is becoming more and more important (Holmes et al., 2015; Chinn & Malhotra, 2002). Measurement uncertainties are an aspect of this data quality (Millar et al., 1994) and research shows that many students have problems when reasoning with measurement uncertainties (see e.g., Hull et al., 2020). The focus of this contribution lies in the comparison of data sets for which measurement uncertainties play a vital role (Kok et al., 2019; Kramer et al., 2017). Students' ability to compare data sets has been investigated before (e.g., Masnick & Morris, 2008; Kapur, 2012; Kung & Linder, 2006). What was not analyzed in these studies, however, is what students actually do when comparing data sets. An analysis of their given justifications (bridging the gap between claim and evidence) could bring this to light, giving more insight into students' understanding about measurement uncertainties. An analysis of justifications has been successfully done by Ludwig et al. (2019). The authors looked at the type of justifications–more rational or more intuitive–students give when supporting or rejecting a hypothesis in the light of anomalous data. Again, the authors did not look at how students actually do data evaluations.

Research Question

The work presented here is part of a PhD project and will focus on the results of the analysis of justifications in a data comparison problem. The research question addressed here is:

RQ How does the quality of students' justification in a data comparison problem change with increased teaching of concepts regarding measurement uncertainties?

Study Design

To answer the research question, a digital learning environment (DLE) was developed, based on the subject matter model by Priemer and Hellwig (2018). The DLE consists of five steps, addressing different conceptual pieces of knowledge (see Tab. 1). The chosen quantification of the uncertainty in step IV is the min-max interval to keep the calculation mathematically simple (Hellwig, 2012, p. 124). Participants are randomly assigned to one of three groups: A, B, or C. Depending on the group, participants will go through a different number of steps.

Step	Concept	Group		
I	Sources of Uncertainty	A	B	C
II	Relevance and Meaning of Measurement Uncertainties	🟩	🟧	🟦
III	Calculation and Meaning of the Mean Value		🟧	🟦
IV	Calculation and Meaning of the Measurement Uncertainty		🟧	🟦
V	Comparing Measurement Results			🟦

Tab. 1 The different steps that make up the DLE and which group goes through which steps

The effect of the DLE is measured in a pre-post design. The pre- and post-test each contain (among other items) a data comparison problem.

Due to the restrictions of the Covid-pandemic, the study was conducted online between April and August 2020. Participation was voluntary and participants worked alone by themselves. A total of 342 participants from grade levels 8 through 12 participated of which $N = 154$ remained that filled in the pre- and post-test and spend more than 5 minutes on the DLE. The average time spend on the DLE was about 60 min, the pre- and post-test each took 10 min.

In the data comparison problem, participants are introduced to a hypothetical skateboard competition in which they play the role of the jury. To participate, each skateboard should have a similar rolling capacity. To determine this, each skateboard is rolled down an incline six times and the rolling times are measured. The same is done for a reference skateboard. The data is shown to the participants (see Tab. 2), and the participants are asked: Based on this data, do you allow the participant's skateboard to the competition? Explain your answer.

Participant Skateboard (s)	Reference Skateboard (s)
1.530	1.548
1.573	1.534
1.522	1.520
1.548	1.571
1.583	1.523
1.538	1.526

Tab. 2 The data that was shown to participants in the data comparison problem

Coding the justifications

Upon analysis, the justifications each showed a similar pattern. Participants base their decision on the data (or not), compare a certain quantity, and check whether a certain criterion is met (or not). For example: *"No. The time the contestant's skateboard takes is on average 1.549 s, the jury skateboard takes 1.537 s. Thus, the contestant's skateboard takes longer."* This justification is based on the data, compares the mean value, and checks whether one mean is larger than the other. Based on this, a coding manual has been developed inductively through several rounds of identification, coding, discussion, and revision. The result is a manual of hierarchical codes for the comparison and criterion, as well as a yes/no (da.yes/da.no) code for whether or not the decision is based on the data. The hierarchy is based on the point and set-paradigms by Lubben et al. (2001) as well as the sophistication and correctness—together *quality*—that can be achieved using this particular comparison or criterion. A sample of 30 justifications has been double coded by the authors, a weighted Cohen's Kappa shows almost perfect agreement: $\kappa_{\text{dat}} = 1.0$, $\kappa_{\text{comp}} = .93$, $\kappa_{\text{crit}} = .99$.

The comparison codes are (from low to high): *Unclear*, compared quantity is unclear; *single measurement*, a comparison of a single measurement; *uncertainties*, a comparison of the measurement uncertainty; *pairs*, a comparison of pairs of measurements; *mean value*, a comparison of the mean values; *deviations*, a new set of deviations is built and evaluated on its own; *sets*, a comparison of the set as a whole, or within and between summarizing quantities; *uncertainty intervals*, a comparison of the uncertainty interval.

The criterion codes are (from low to high): *Unclear*, the deciding criterion is unclear, absent, or irrelevant; *duplicates*, the deciding criterion is the presence (or absence) of duplicates; *larger*, a quantity is larger/smaller than the other; *counts*, the number of occurrences of something; *closeness*, how "close" one quantity is to the other; *overlap*, two intervals/ranges overlap (or not).

Results

The results of the pre-test show no difference in distributions of comparison or criterion code between groups A, B, and C ($p > .1$), indicating homogeneity between groups.

Figure 1 shows the results of the post-test. The results of the pre-test are combined and added for reference in gray. In contrast to the pre-test, the post-test code distributions are significantly different between groups A, B, and C with large effect sizes ($\chi^2(14) = 53.36$, $p < .001$, $w = .59$, $\chi^2(12) = 47.47$, $p < .001$, $w = .56$ resp.). In general, what can be seen is an increasing shift towards higher quality codes between groups, as well as a stepwise increase in both the interval comparison and the overlap criterion codes between groups.

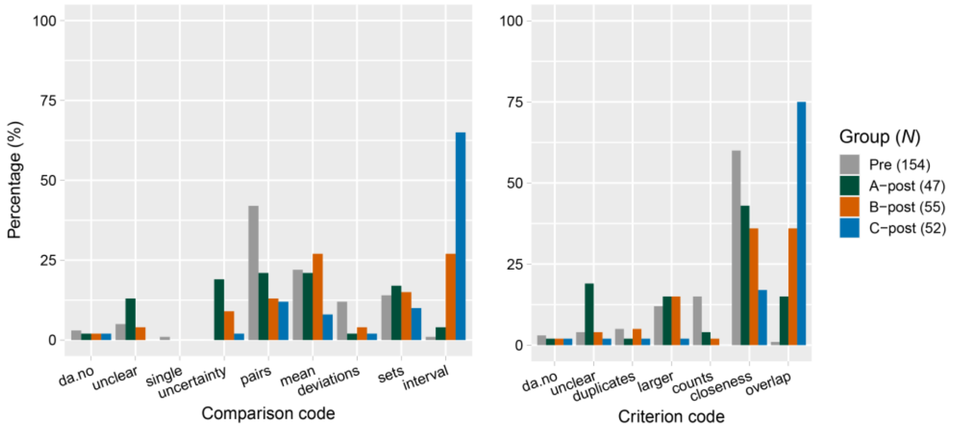

Fig. 1 The percentage distribution of the comparison and criterion codes of the (combined) pre-test (gray) and the post-test (A, green; B, orange; C, blue). "da.no" indicates that the justification was not based on the data. Hierarchy of codes is from left to right.

The differences between the comparison code distributions of each group's individual pre- and post-test are all highly significant with large effect sizes.
In group A, the pairwise comparison has decreased and the codes are distributed more or less evenly over all other codes (13–22 %), including the unclear code. For group B the pairwise comparison decreased even further. The dominant codes are the mean and interval (both 27 %). In group C nearly all codes have decreased in favor of the interval code (65 %).

The differences between the criterion code distributions of each group's pre- and post-test are all significant, the effect sizes are medium for group A and large for B and C.
The dominant criterion code for group A is closeness (60 %). Remarkable is here, again, the increase in the unclear code. In group B, closeness decreased further but remains, together with the overlap code, the dominant criterion code (both 36 %). In group C, the closeness criterion decreases even further and the overlap criterion is by far the dominant code (75 %).

When looking at the influence of grade level on the distributions of codes, no significant differences in the code distributions between grade levels can be found in the pre-test, (p > .1) for both the comparison and the criterion code distributions. In the post-test, the same holds for the criterion code. For the comparison code this difference is just significant with a medium to strong effect size, (p = .04, w = .47). However, no gradual shift towards higher quality comparison codes can be seen with increasing grade level but rather, the difference appears erratic without any meaningful pattern.

Conclusion

First of all, the coding of the justifications allows for a quick overview of what students actually do when they are comparing data sets. The hierarchy of the codes gives a measure of students' ability to compare data sets.
From the stepwise increase in interval comparison and overlap criterion codes between groups A, B, and C, we conclude that increased teaching about measurement uncertainties leads to higher quality justifications (RQ).
Lastly, the quality of these justifications in both the pre- and the post-test is the same for grade levels 8 through 12. This means that the topic of measurement uncertainties can be successfully introduced as early as 8th grade.

Literature

Chinn, C. A. and Malhotra, B. A. (2002). Epistemologically Authentic Inquiry in Schools: A Theoretical Framework for Evaluating Inquiry Tasks. *Science Education, 86*(2):175–218 https://doi.org/10.1002/sce.10001

Hellwig, J. (2012). *Messunsicherheiten verstehen: Entwicklung eines normativen Sachstrukturmodells am Beispiel des Unterrichtsfaches Physik* Doctoral Thesis, Ruhr-Universität

Holmes, N. G., Wieman, C. E., and Bonn, D. A. (2015). Teaching critical thinking. *Proceedings of the National Academy of Sciences, 112*(36):11199–11204 https://doi.org/10.1073/pnas.1505329112

Hull, M. M., Jansky, A., and Hopf, M. (2020). Probability-related naïve ideas across physics topics. *Studies in Science Education*, pages 1–39 https://doi.org/10.1080/03057267.2020.1757244

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science, 40*(4):651–672 https://doi.org/10.1007/s11251-012-9209-6

Kok, K., Priemer, B., Musold, W., and Masnick, A. (2019). Students' conclusions from measurement data: The more decimal places, the better? *Physical Review Physics Education Research, 15*(1):010103 https://doi.org/10.1103/PhysRevPhysEducRes.15.010103

Kramer, R. S. S., Telfer, C. G. R., and Towler, A. (2017). Visual Comparison of Two Data Sets: Do People Use the Means and the Variability? *Journal of Numerical Cognition, 3*(1):97–111 https://doi.org/10.5964/jnc.v3i1.100

Kung, R. L. and Linder, C. (2006). University students' ideas about data processing and data comparison in a physics laboratory course. *Nordic Studies in Science Education, 2*(2):40–53 https://doi.org/10.5617/nordina.423

Lubben, F., Campbell, B., Buffler, A., and Allie, S. (2001). Point and set reasoning in practical science measurement by entering university freshmen. *Science Education, 85*(4):311–327 https://doi.org/10.1002/sce.1012

Ludwig, T., Priemer, B., and Lewalter, D. (2019). Assessing Secondary School Students' Justifications for Supporting or Rejecting a Scientific Hypothesis in the Physics Lab. *Research in Science Education*, *51*(3):1–26 https://doi.org/10.1007/s11165-019-09862-4

Masnick, A. M. and Morris, B. J. (2008). Investigating the Development of Data Evaluation: The Role of Data Characteristics. *Child Development, 79*(4):1032–1048 https://doi.org/10.1111/j.1467-8624.2008.01174.x

Millar, R., Lubben, F., Got, R., and Duggan, S. (1994). Investigating in the school science laboratory: Conceptual and procedural knowledge and their influence on performance. *Research Papers in Education, 9*(2):207–248 https://doi.org/10.1080/0267152940090205

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. https://doi.org/10.1007/s10763-016-9768-0