Exploring Student Sensemaking When Engaging in Data Cleaning

COLLEGE OF EDUCATION | THE UNIVERSITY OF UTAH
Learning and Cognition

Adrian Adams[1], Lauren Barth-Cohen[1,2]
[1]Department of Educational Psychology, University of Utah
[2]Department of Physics and Astronomy, University of Utah

Presenter Contact:
AL.Adams@utah.edu

Background

- Scientists evaluate data regularly in preparation for analysis (Roth, 2013; Jin & Hokayem, 2021)
- Instructors can teach data literacy by engaging students with messy data (Kjelvik & Schultheis, 2019)
- Data Cleaning: Identifying and mitigating the effects of anomalous data
- Students do engage in data cleaning (Cummiskey, Kuiper, & Sturdivant, 2012; Johansen, & Christiansen, 2020), but little research on how specifically they do so

Framework

- Epistemological Resources (Hammer & Elby, 2002)
  - Beliefs about acquiring and using knowledge should impact working with data
- Epistemic Games (Collins & Ferguson, 1993)
  - Actions one can take to understand information

Research Question

What epistemological resources are students using, and what epistemic games are students playing, when they encounter anomalous data?

Methods

- Multiple interviews with 6 students enrolled in Physics for the Life Sciences Lab course
  - 4 biology, 1 kinesiology, 1 communications major
  - 5 self-identified as female, 1 as male
- Students encouraged to bring data from lab
- Students interacted with scenarios designed to mimic lab data
- Computer screen and audio recorded
- Transcripts coded to capture student use of epistemological resources (what fuels student interpretation and decisions) and epistemic games (how students work with the data to sensemake)

Future Steps

Continue analysis to determine if there is change in resource use over time

Acknowledgments

This material is based upon work supported by the National Science Foundation Grant No. 1938721.

Results

- Epistemic Games Identified
  - *Using patterns to make sense of data* (Rodriguez, Bain, and Towns, 2020): Looking for and identifying patterns in qualitative (e.g., video) or quantitative (e.g., coordinates in a spreadsheet) data without manipulation
  - *Using trends to make sense of data* (Collins & Ferguson, 1993): Representing data in a graph or plot to understand behavior of data
- Epistemological resources impacted student decisions during and after epistemic games
- Students utilize beliefs and intuitions to interpret and manipulate their data
  - Prior experiences in lab settings influence these beliefs/intuitions
- **Figures 1 & 2: Phrases in italics were coded as the resource in {curly brackets} at the end of the line**

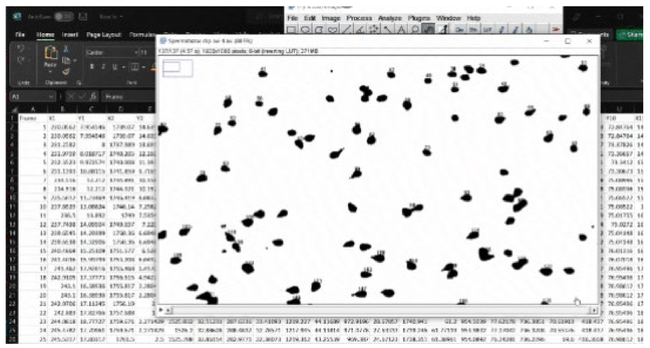

Figure 1. Example of Using Patterns To Make Sense Of Qualitative Data

(Student is viewing a video of sperm cells swimming. The cells have been auto-tracked via ImageJ software, producing x and y coordinates and are labeled with object numbers. Student is watching the processed video.)

Julia: They pass off the numbers [object labels], when there like, near each other. Like one will have the number, like this one, like after it passes it, that one doesn't have the number any more.

Interviewer: Yeah.

Julia: I almost wonder if it would be better to, *just select from here? like one's that don't have that problem to use*, rather than finding every single one where it does that, cuz it happens a lot. {Select data to avoid anomalies}

Interviewer: Yeah I guess the question I might ask is, is do you want to try to like salvage any data from the ones that do end up kind of crisscrossing?

Julia: Uh, not really, I mean if this was like a lab report, I probably would, but I don't know, *for the sake of the time, I wouldn't aim for it*. {Efficiency of action impacts doing that action}

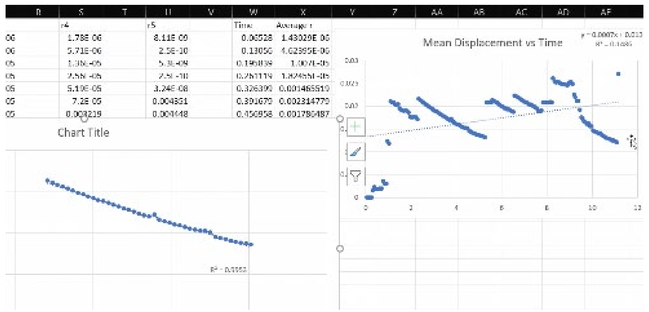

Figure 2. Example of Using Trends to Make Sense of Data

(Student is determining how to acquire a slope from the plot of mean displacement over time to use in a diffusion coefficient calculation.)

Kristen: I think it'd-*averaging [across lines] would do better like, hm, make for more precise number [slope], more accurate*. {Use multiple trials to generate conclusions}

Kristen: But I think like, personally for me, *if I was going through lab, and wanted to save time [laugh], I'd just pick one and hope it's constant for the rest of the trial*... {Efficiency of action impacts doing that action}

Interviewer:…So if you had to pick like just one segment to use of this data.

Kristen: I definitely want this one [graphed on the left], cuz I think it's the most [garbled] I think it has the best fit of them. The next one would be the very right one [on the right plot], *cuz it's kind of a curve, but I feel like a trendline could still fit pretty well*. {Relative impact of anomaly matters when dealing with anomaly}

Interviewer: So what makes, what about the data gives you I guess, indicators of like a good chunk of data to you?

Kristen: *I think if it's pretty consistent, [garbled] and it has like the most steady results*…{Data should fit a trendline relatively well}

Interviewer: So is amount of data important in your decision or is it more like the smoothness?

Kristen: I think amount, *because the more that like it fits what you're graphing, gives you a better representation of what your result is*. {select data to be representative}

References

Collins, A., & Ferguson, W. (1993). Educational *psychologist,* 28(1), 25-42.
Cummiskey, K., Kuiper, S., & Sturdivant, R. (2012).. *Frontiers in psychology*, 3, 354.
Hammer, D., & Elby, A. (2002). *The psychology of beliefs about knowledge and knowing*, 169190.
Jin, H., Hokayem, H., & Cisterna, D. (2021). *Research in Science & Technological Education*, DOI: 10.1080/02635143.2021.1902977
Johansen, M. W., & Christiansen, F. V. (2020). *Science and Engineering Ethics*, 26(2), 1107-1128.
Rodriguez, J. M. G., Bain, K., & Towns, M. H. (2020). *International Journal of Research in Undergraduate Mathematics Education,* 6(2), 279-301.
Roth, W. M. (2013). *Research in Science Education*, 43(4), 1617-1644.