International Society of the Learning Sciences

How do middle school students think about climate change?

Sarah Bichler, University of California, Berkeley, sbichler@berkeley.edu,
Allison Bradford, University of California, Berkeley, allison_bradford@berkeley.edu,
Brian Riordan, ETS, briordan@ets.org,
Marcia C. Linn, University of California, Berkeley, mclinn@berkeley.edu

We use natural language processing (NLP) to train an automated scoring model to assess students' reasoning on how to slow climate change. We use the insights from scoring over 1000 explanations to design a knowledge integration intervention and test it in three classrooms. The intervention supported students to distinguish relevant evidence, improving connections between ideas in a revised explanation. We discuss next steps for using the NLP model to support teachers and students in classrooms.

Objective

This research connects two studies to characterize students' understanding of factors impacting climate change and improve student ability to evaluate actions that slow climate change.

Embedded assessment "Group Effort"

We designed the Group Effort item to engage students in reasoning about ways to slow climate change. Students use an online data tool with rates of electricity usage for several devices from a local gas and electric company to evaluate plans for reducing electricity usage proposed by three fictitious students. Key to the analysis is that each device uses electricity at a different rate. Students record the pounds of CO_2 saved for each plan to compare, evaluate, and explain what the three friends could do to increase their impact.

Knowledge integration guidance

The knowledge integration (KI) framework outlines four key processes to help students develop coherent understanding of science phenomena: *elicit* student ideas so they can be compared and refined, *discover* new ideas by exploring models, *distinguish* between initial and new ideas, and *reflect to connect* their ideas (Linn & Eylon, 2011). Curriculum designs using the KI principles engage students in multiple cycles through all four KI processes, providing students with opportunities to continuously integrate their new ideas with their prior understanding (Gerard et al., 2020). Student performance on embedded assessments might reveal that they still need support to fully integrate their ideas. We argue that, at this point, students benefit from guidance to distinguish between alternative perspectives or conflicting evidence.

Study 1: Students' ideas and NLP model

Participants, data source, and scoring approach

Seventeen middle and high school teachers from 8 diverse schools administered an online science assessment at the end of the school year, including the Group Effort question. We analyzed 1146 student responses.

We developed a scoring rubric (1-3) to assess students' understanding of the mechanism (connecting energy use to CO_2 emissions and human impact on climate change) and data practice (using evidence to evaluate plans). A knowledge integration score (1-5) indicates whether students connected ideas within and/or across dimensions. Two researchers independently coded 70 responses, iteratively refining the rubrics until reaching Cohen's Kappa of > .85 on each coded dimension. Then each researcher coded 50% of the entire data set.

NLP model building & results

We used a state-of-the-art neural network scoring model architecture that leverages a pre-trained transformer language model "fine-tuned" for the KI scoring task. These models have been pre-trained to identify words that have been masked from the input and to predict whether one sentence follows another. The word representations the model learns to do these tasks and the self-attention based transformer architecture, yields word representations that are useful across many NLP tasks. We used a model based on the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2019). This model architecture, adapted for the KI scoring task, has achieved state-of-the-art performance on the ASAP Short Answer Grading dataset (Steimel & Riordan, 2020). The models were trained to predict an ordinal score from each response's text. Since KI is concerned with the content of the response, the models did not consider grammatical or usage errors in scoring.

Human-machine agreement was good with weighted Kappa of .79 for the mechanism, .88 for data practice and .89 for KI. By analyzing students' responses, we learned that students often used time on a device as evidence to judge the effectiveness of the solutions rather than kWh of energy use, indicating students need support to distinguish between time a device is used and how much energy per hour each device uses.

Study 2: KI guidance intervention

To help students distinguish between relevant evidence, we developed a KI guidance intervention. Students use a data tool to analyze a TV, video console, and desktop computer. They reduce the use of each device from 2 to 1 hour, enter the evidence in a table and compare how much CO_2 is saved. Students then rank each device according to how many kWh and pounds of CO_2 are saved. Students then revise their Group Effort explanation.

Participants, data source, analytical approach, and statistical analysis

Three middle school science teachers from two schools and their 397 6th grade students participated in the study. The Group Effort item is embedded in an online unit on global climate change featuring interactive models and activities (https://wise.berkeley.edu/). A total of 345 students completed the activity during remote instruction.

We used the NLP model to score the initial and revised responses to the Group Effort item which are logged in the learning environment. We used repeated measures ANOVA to test whether students' understanding improved from before to after the KI intervention. We used the open-source statistical software jamovi, set an alpha level of 5% and applied Bonferroni correction ($\alpha/n = .05/3 = .02$) to control for multiple testing.

Effectiveness of the KI intervention

We used the initial KI score as measurement point 1 and the revised KI score as measurement point 2 and included the teacher as a between subjects factor to test if there was a teacher effect. Students' KI scores increased from initial ($M = 2.79$, $SE = 0.05$) to revised explanation ($M = 3.23$, $SE = 0.06$), $F(1, 342) = 90.42$, $p < .001$, $\eta^2 = 0.04$. A non-significant interaction effect between time and teacher indicates that students' KI scores did not increase at different rates across teachers ($F(2, 342) = 1.36$, $p = .257$, $\eta^2 = 0.001$).

To unpack whether the change in KI is due to students connecting more ideas to explain the mechanism or use data as evidence, we used repeated measures ANOVA. Students connected more ideas about the mechanism in the revised DCI ($M = 1.42$, $SE = 0.04$) than in the initial ($M = 1.24$, $SE = 0.03$) explanation $F(1, 342) = 43.11$, $p < .001$, $\eta^2 = 0.02$. They also connected more data as evidence: initial $M = 1.50$, $SE = 0.04$ and revised $M = 1.77$, $SE = 0.04$ explanation, $F(1, 342) = 62.44$, $p < .001$, $\eta^2 = 0.03$.

Discussion and conclusion

The NLP scoring model can be used in future classrooms to provide teachers with information about their students' reasoning or to adaptively guide students (Gerard et al., 2020). The KI intervention targeted towards sorting ideas not only helped students to engage in the sophisticated practice of distinguishing between time versus kWh as evidence, but also helped students integrate this data practice with a mechanistic understanding as they connected their ideas about electricity usage, CO2 emissions, and human impact on climate.

References

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1 (long and short papers)*, 4171–4186.

Gerard, E., Wiley, K., Bradford, A., King Chen, J., Lim-Breitbart, J., & Linn, M. (2020). Impact of a Teacher Action Planner that Captures Student Ideas on Teacher Customization Decisions. In M. Gresalfi & I. Seidel Horn (Eds.), *14th International Conference of the Learning Sciences* (Vol. 1, pp. 2077–2084).

Linn, M. C., & Eylon, B.-S. (2011). *Science learning and instruction: Taking advantage of technology to promote knowledge integration*. Routledge.

Steimel, K., & Riordan, B. (2020). Towards instance-based content scoring with pre-trained transformer models. *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, 34*.

Acknowledgments

We received funding from grant 1813713 STRIDES: Supporting Teachers in Responsive Instruction for Developing Expertise in Science from the National Science Foundation. We thank the participating teachers and students for collaborating with us.