International Society of the Learning Sciences

Adaptive Dialog to Support Student Understanding of Climate Change Mechanism and Who is Most Impacted

Allison Bradford, University of California, Berkeley, allison_bradford@berkeley.edu
Weiying Li, University of California, Berkeley, weiyingli@berkeley.edu
Brian Riordan, ETS, briordan@ets.org
Kenneth Steimel, ETS, ksteimel@ets.org
Marcia C. Linn, University of California, Berkeley, mclinn@berkeley.edu

Abstract: To support ninth graders to take advantage of the ideas, intuitions, and experiences that contribute to their understanding of climate change, we designed an NLP-based adaptive dialog and tested it in a week-long unit exploring Urban Heat Islands. The dialog's guidance prompts were designed to elicit students' ideas about climate change mechanisms. Students interacted with the adaptive dialog twice during the unit. We scored their initial and revised explanations (after engaging with the adaptive dialog each time) using a Knowledge Integration (KI) rubric. A repeated measures mixed ANOVA revealed that students who initially expressed descriptive ideas often rooted in experience made significantly greater gains during the dialog than those who initially expressed mechanistic ideas. The dialog supported all students to broaden the ideas they considered when exploring climate change. Further, a paired t-test revealed that students made overall gains in KI from pretest to posttest (d=.76).

Introduction

The ideas and experiences students express while reasoning about scientific phenomena are productive and powerful resources for developing integrated understanding (diSessa, 2006; Linn, 2006; Hammer, 2000) especially when they are taken up and built upon by teachers (Fulberg & Silseth, 2022). Students' funds of knowledge (González, Moll, & Amanti, 2005) are grounded in their family culture, daily routines, and community interactions and inform their sensemaking in science classrooms. The initial ideas students hold are often a combination of vague, descriptive, and mechanistic ideas (diSessa, 2006). Research shows that when instruction neglects the ideas students hold while introducing new ideas, students develop a fragmented understanding by either dismissing their own ideas within the school context in favor of what is presented during instruction (Carlone, 2015) or holding both perspectives rather than evaluating and making connections across ideas (diSessa, 2006).

We explore the impact of an adaptive dialog that responds to specific ideas detected in student explanations, mirroring an effective teacher responding to and building upon students' ideas. In particular, this study examines the effect of the adaptive dialog on the learning gains of students who initially express primarily descriptive ideas such as those that are vague and not yet clearly linked to an explanatory mechanism as compared to students who initially express primarily mechanistic ideas. Descriptive ideas are less likely to be taken up and built upon during class interactions (Bang & Medin, 2010) than mechanistic, school-based ideas, which the teacher might more easily recognize as relevant to classroom discussion. We investigate the role of an adaptive dialog that recognizes and builds upon students' descriptive ideas.

Advanced natural language processing (NLP) techniques enable detection of individual ideas in each student's written explanation (Riordan et al., 2022) and have the potential to be embedded within science curricula to help students value and build on their initial ideas. However, a recent review of adaptive dialogs in education, also called chatbots, found that few studies were carried out in K12 classrooms. Further, these studies had negligible learning gains (Wollny et al., 2021). Wollny et al. suggested that the lack of impact might reflect that "chatbot development in education is still driven by technology, rather than having a clear pedagogical focus of improving and supporting learning" (p.13).

We draw on the Knowledge Integration pedagogical framework (KI; Linn & Eylon, 2011) with the aim of designing effective adaptive dialog that builds on students' funds of knowledge to promote development of an integrated understanding of climate change. We investigate the impact of dialogs with guidance that is adaptive to students' distinct initial ideas. KI advocates respecting and building on the diverse ideas each student brings to the classroom. KI recognizes that learners hold varied ideas that reflect their lived experiences. The KI pedagogical model supports students to develop coherent understanding by: *eliciting* and valuing the full range of students' ideas; providing opportunities for students to *discover* evidence in a variety of relevant and meaningful contexts; using evidence to *distinguish* among existing ideas and new ideas; and guiding students to *make connections* among their ideas to form an explanation or argument (Linn & Eylon, 2011).

KI informed the design of both the adaptive dialog and the unit under study. The adaptive dialog is driven by an NLP idea detection model that is trained to identify a wide range of ideas, including ideas rooted in experience, vague ideas, and nonnormative ideas. Adaptive guidance prompts are assigned based on the ideas detected and are designed to elicit further details and reasoning about the ideas students initially expressed. The unit in this study aims to *elicit* students' observations of the causes and impacts of climate change and Urban Heat Islands. It guides students to *discover* new evidence and perspectives. It prompts students to *distinguish* among possible mechanisms that explain climate change and to understand who in their community is most impacted by it. And, it supports students to *make connection*s to explain causes and formulate solutions. In this study we investigate how the design of the unit, including the adaptive dialog, supports students to build on their own ideas to understand the mechanism causing climate change as well as who is most impacted by the dialog.

Methods

Participants

The study was implemented in a suburb of a large Western city in the United States with two ninth grade biology teachers and all of their students. The school is racially diverse, has 26% of students eligible for free or reduced-price lunch and 5% of students are emergent multilingual learners. Only students who completed pretest, posttest, and two full interactions with the adaptive dialogs were included in the analysis (N=70).

Curriculum

David and Mary both taught the Global Climate Change and Urban Heat Islands (UHI) unit to their 9th grade biology students. This unit features interactive models, data visualizations, and mapping activities (Bradford et al., 2022). Students first explore how energy from the Sun is transferred and transformed to heat the Earth. Their ideas and experiences are elicited through activities that ask them to make predictions. They use interactive models to discover ideas about the natural greenhouse effect and then investigate how human activities can amplify the greenhouse effect leading to anthropogenic climate change. Students are then prompted to consider the impact climate change has on people. They are introduced to several youth climate activists from their community and are asked to consider whether marginalized communities experience greater effects of climate change than dominant communities. Students explore the UHI phenomenon through investigations of how different surfaces are heated by the sun. Students deepen their understanding by distinguishing how historical, racist policies, like redlining, contribute to some areas becoming UHIs. Students compare historical redlining maps to present day temperature maps to gather evidence of the impact of redlining.

Pre and posttest assessment

Pretests and posttests were administered to students before and after interacting with the curriculum. Two items determined the pretest and posttest scores used in analysis. The *Coal* item prompted students to explain how increased levels of carbon dioxide in the atmosphere might impact the climate. This item examines students' conceptual understanding of the mechanisms of climate change. The *Impacts* item prompted students to explain how the effects of climate change could impact people and whether or not all people are impacted by these effects in the same way. This item measures connections students make between the causes of climate change and the historical policies that have shaped who experiences the impacts of climate change, such as the effects of UHI. Validated Knowledge Integration rubrics were used to automatically score student responses to each item (Liu et al., 2008). KI scores indicate the degree of integration among normative science ideas that students have in their explanations, without penalizing for vague or non normative ideas. The KI scoring rubric is on a scale of 1-5. The scoring rewards students for connecting their ideas to evidence and does not penalize students for expressing vague or non-normative ideas. A score of 1 or 2 indicates the student has not yet connected their ideas to an explanatory mechanism, while a score of three or higher reflects that students have connected at least one mechanistic idea.

Adaptive dialog idea scoring

The interactive, adaptive dialog occurred on the *Car* item (Figure 1) at two points in the unit: early in the unit after completing a lesson about the natural greenhouse effect, and right before taking the posttest. The adaptive dialog was designed to elicit and build upon student explanations about how the temperature inside a car would compare to the outside temperature on a cold, sunny day. This is an analogy for the greenhouse effect mechanism that drives climate change and leads to rising global temperatures. It asks students to connect the same ideas students use to explain the pre/posttest item *Coal.*

To develop the adaptive dialog, we built a *multi-label* NLP model for idea detection. The NLP model used a *token classification* approach for idea detection (Riordan et al., 2020; Schulz et al., 2018). The model was trained to predict an idea category for each word token in the student response data. Consecutive words with the same predicted idea were treated as an idea *span.* Since ideas can overlap, a given word token can receive more than one idea category prediction. The multi-label idea detection model was trained and validated with 10-fold cross validation for hyperparameter tuning and evaluation. To prepare data to train the model, we first created an idea rubric (Table 1) that enumerates the ideas typically expressed in response to the item, including ideas rooted in personal experience, vague ideas, and non-normative ideas in addition to the ideas that comprise the mechanism targeted by the item (for details see Gerard et al., 2022). We generated a set of 24 ideas for the *Car* item which included nine mechanistic ideas, seven vague or personal experience-based ideas, and eight nonnormative ideas. We then human-coded 793 student explanations by annotating spans of text and labeling them with the corresponding idea. The human-coded student explanations were collected at schools serving similar populations to the school in the present study. Prior to use in this study, the idea detection model was evaluated on word-level micro-averaged F-score (harmonic mean of precision and recall, weighted by frequency of idea category). The overall F-score achieved by the model was .592, in line with similar models developed for challenging idea and reasoning detection tasks (Schulz et al., 2018; 2019).

We also build a KI scoring model (using a KI scoring rubric) for the initial and revised responses to the *Car* item, using 1093 human-coded student explanations. The model achieved a Quadratic Weighted Kappa (QWK) of .889. This indicates a high level of agreement between model and human scoring, exceeding the standard threshold of 0.7 for automated scoring deployment (Williamson et al., 2012).

Table 1

Sample from Car Item Idea Rubric and Adaptive Guidance. The full rubric includes 24 ideas across all KI levels and includes both descriptive and mechanistic ideas.

Item prompt: *On a cold winter day, Akbar is walking to his car that is parked in the sun. His car has not been driven for one week. Does the air inside the car feel colder, hotter or exactly the same as the outside air? Explain.*

#	Idea	KI	Assigned Adaptive Eliciting Guidance in Dialog
2	The car feels colder than outside (DI)	2	Interesting ideas! Can you tell me about how you feel if you are sitting inside a car and the sunlight shines on you through the window?
5	Heat (or cold) is conducted between the car and the surrounding area (DI)	2	You're right that heat energy is conducted between objects that touch each other. Since the Sun is far away, how can its energy warm the car?
8	The sun warms the car directly (DI)	2	Interesting idea! How does the sun make the car warmer?
15	Solar radiation is transformed into heat energy (MI)	3	I think you're saying that energy from the sun transforms to heat energy in the car. What happens to that heat energy?
18	Heat or infrared energy is trapped inside the car (MI)	3	Can you tell me more about why heat gets trapped in the car but solar radiation doesn't?

Within the adaptive dialog, the idea detection model was used to detect the ideas present in each response the students provided. After their initial response, students received an adaptive guidance prompt based on the ideas detected (Figure 1). Drawing on the KI framework, prompts were designed to elicit more of the students' thinking about the idea they had expressed. Because multiple ideas could be detected in each student response, we developed a prioritization order for which ideas to guide. We opted to guide detected mechanistic ideas before vague or non-normative ideas. In the dialog interface, students responded to the adaptive prompt and then were given a second, non-adaptive prompt that asked them to share an idea they were unsure about. The second prompt was non-adaptive because we trained the idea detection model on full explanations and had not yet determined how it would perform at identifying student ideas in their responses to the first adaptive prompt. After responding to the non-adaptive prompt, students were asked to revise their explanations.

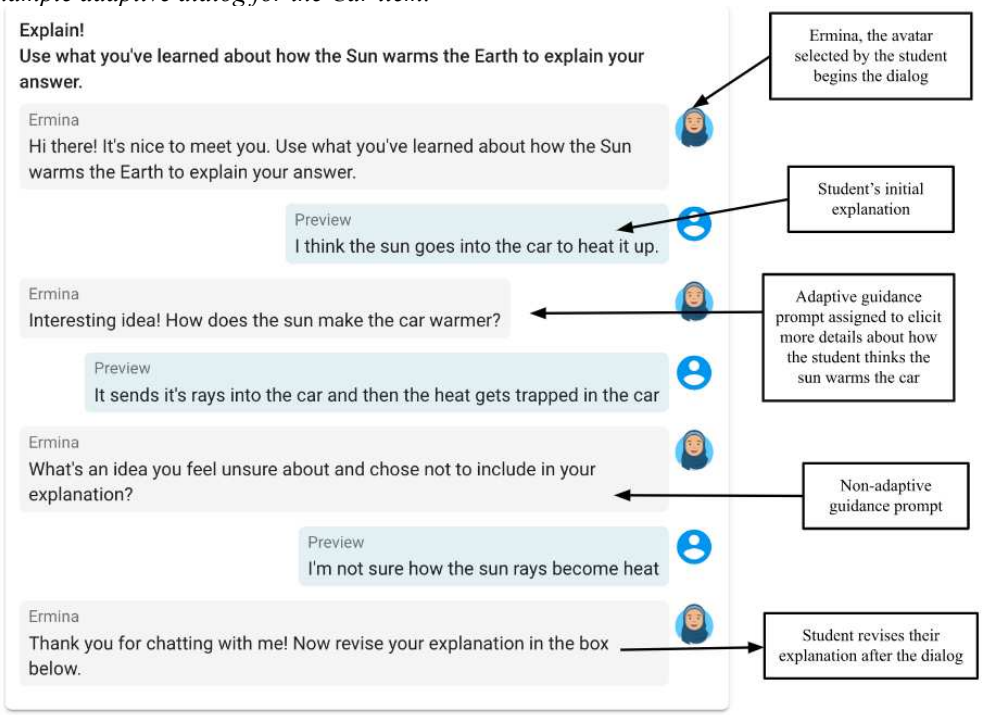
Analysis approach

Descriptive and mechanistic student idea categories

When responding to the Car item, students express a mix of descriptive ideas rooted in everyday experience and mechanistic ideas that are often connected to evidence that they have explored throughout the unit. Because we were interested in how well the unit and dialog supported students to connect their descriptive ideas to an explanatory mechanism, we categorized students based on the kinds of ideas they initially expressed. To assign students to the categories, we considered their performance on both pretest items and on their initial response to the Dialog Item (Car). For each of these items, a score of 3 or more must have at least 1 mechanistic idea connected to evidence, whereas a score of 1 or 2 indicates the use of descriptive or vague ideas not yet connected to a mechanism. We assigned students who had scores of 1 or 2 and not more than one 3 across these items to the Descriptive Ideas (DI) category. We assigned students with at least two scores of 3 or more on each item to the Mechanistic Ideas (MI) category. Thus, students in the DI category started with not more than one mechanistic response across the items. Students in the MI category started with two or more mechanistic ideas. There were 25 students in the DI category and 45 in the MI category.

Figure 1

Example adaptive dialog for the Car item.



Progress across dialog items

To analyze the impact of the adaptive dialog, we examined the frequency of ideas detected during each round of guidance in each dialog as well as in the revised explanations. We computed the frequency and types of ideas added and pruned between initial and revised explanations for each interaction with the dialog, comparing patterns for DI and MI students. We also use a repeated measures mixed ANOVA to examine the change in students' KI score over *time* (four time points: initial explanation in the first engagement with the dialog, revised explanation after the first dialog, the initial explanation in the second dialog, and final revised response after the second dialog) with student idea category (DI and MI) as a between-subjects factor.

Progress of DI and MI students from pretest to posttest

We assessed students' overall change in understanding by conducting a paired t-test on students' scores on both pretest and posttest items. We computed the gain students made on each item by taking the difference between posttest and pretest scores. We then used a two-sample t-test to investigate whether the gains on either item varied by student idea category, to account for the alignment of the dialog with the *Coal* item and not the *Impacts* item.

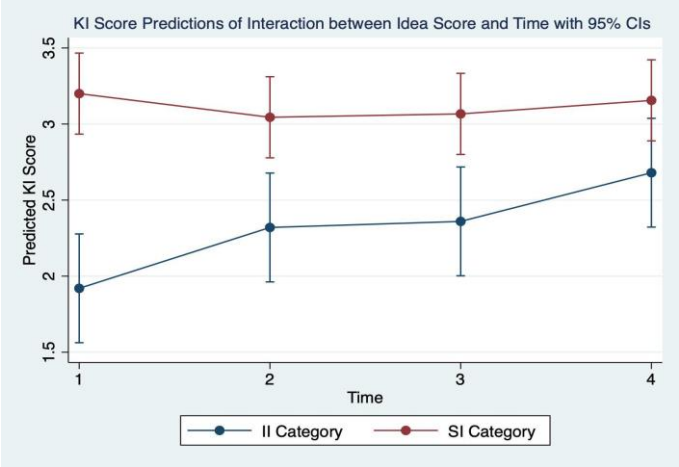
Results and discussion

Progress across dialog items

Across the Dialog items, a repeated measures mixed ANOVA using the Greenhouse-Geisser correction revealed a main effect for both *time* ($F(3, 68)=3.41$, $p=0.022$, $\eta^2=0.048$) and student *Idea Category* ($F(1, 68)=19.8$, $p<0.001$, $\eta^2=0.226$) along with a significant interaction between the factors ($F(3, 68)=4.51$, $p=0.005$, $\eta^2=0.062$). Post-hoc analysis of estimations revealed that DI students made significant gains in KI score at all time points when compared to their explanation at time point 1. The greatest significant difference in KI score for DI students ($t(68)=4.17$, $p<0.001$) was between initial explanation ($M=1.92$, $SD=0.75$) and final revision after the second dialog ($M=2.68$, $SD=0.80$). The KI score difference for MI students was not significant between any time points (Figure 2). Thus, the two interactions with the adaptive dialog on the Car item were more beneficial for DI students than for MI students.

Figure 2

Estimated KI Scores on the Car item at each time point by student Idea Category with 95% CIs



Idea changes during dialog interactions

For both dialogs, students generated ideas in multiple categories (Table 2). We analyzed the overall impact of the adaptive dialog as well as the specific ideas students contributed. We also looked at the ideas generated by students who started with DI and MI ideas.

Advantage of adaptive guidance.

For both dialogs, adaptive guidance elicited more ideas (164) than non-adaptive guidance (52). Specifically, adaptive guidance elicited more mechanistic (99 versus 21) and vague ideas (53 versus 14) than the non-adaptive guidance. Thus, tailoring the first guidance prompt to students' initially expressed ideas motivated them to generate more ideas than did the second, non-adaptive prompt, consistent with prior research showing the value of building on students' prior knowledge and experiences (Zacharia et al., 2015).

Table 2

Frequency of ideas elicited at each round during both interactions with the adaptive dialog. Ideas from the rubric are characterized as mechanistic, vague, or nonnormative. Ideas detected were summed across categories. N=70

Dialog	Round	Mechanistic	Vague	Nonnormative	Total
First Dialog	Initial	89	18	14	116
	Adaptive Guidance	49	26	6	81
	Non-adaptive Guidance	14	9	8	31
	Final Revision	81 (62%)	28 (21%)	22 (17%)	131
Second Dialog	Initial	78	24	16	118
	Adaptive Guidance	50	27	6	83
	Non-adaptive Guidance	7	5	9	21
	Final Revision	103 (65%)	36 (23%)	19 (12%)	158

In both opportunities to engage with the adaptive dialog (Table 2), students in the DI category added more ideas on average from initial to revised explanation than students in the MI category (M=1.08 vs M=0.87 and M=1.2 vs M=1.02). In the first dialog, a greater proportion of these added ideas were mechanistic for MI students (43.9%) than for DI students (33.3%). However, in the second dialog, at least half of the ideas added were related to the mechanism for both groups of students. Specifically, the most frequently added idea during the second dialog was idea 18, the idea that "the car traps heat energy." This is a key part of the mechanism. In both dialogs, students in the MI group pruned more ideas on average than those in the DI category. Additionally, a higher proportion of the pruned ideas were mechanistic for the MI category than for the DI category (Table 3).

Specific changes in ideas

The differences in the frequency and types of ideas that students expressed can explain some of the differences in change in KI score for students in the DI and MI groups. Students in the DI category tended to start with vague or non normative ideas and then integrate mechanistic ideas generated during the dialog into their revised explanations. Students in the MI group tended to start with one or more normative ideas in their initial explanations. The dialog served to elicit additional normative ideas from many of the MI students, however, these students tended to focus only on the new idea in their revised explanations rather than integrating them with their initial ideas.

Table 3

Mean frequency of all ideas (and of mechanistic ideas) that were added and pruned between initial explanation and final revision following interactions with the adaptive dialog.

Dialog	PK Level	Ideas Added/Student	Percent Mechanistic Added	Ideas Pruned/Student	Percent Pruned	Mechanistic
First Dialog	DI	1.08	33.3%	0.48	41.7%	
	MI	0.87	43.9%	0.82	59.5%	
Second Dialog	DI	1.2	53.3%	0.52	46.2%	
	MI	1.02	50%	0.88	59.1%	

Typical dialog exchanges for DI and MI students were selected from the second interaction with the dialog (Table 4). The DI student initially explains that inside the car is hotter because the sun is "looking right inside" the car. The idea detection model recognizes idea 8, a vague, descriptive idea about how the sun directly warms the car. Based on detection of this idea, the avatar offers an adaptive guidance prompt that asks the student to explain more about how the Sun makes the inside of the car warmer than the outside. The DI student responds to the prompt with idea 18, the mechanistic idea that the car can trap heat energy. In their final revision, the DI student incorporates the idea about the trapping mechanism. The MI student initially explains that solar (electromagnetic) radiation is emitted by the Sun and that some radiation is reflected while some is absorbed. Their explanation also indicates uncertainty about the distinction between solar radiation and infrared radiation. The idea model detects the presence of two mechanistic ideas: idea 13, that radiation comes from the sun, and idea 14, that some radiation is reflected and some is absorbed. Based on the presence of idea 14, the avatar offers an adaptive prompt that elicits additional ideas from the student, namely what happens to the radiation once it has been absorbed. The MI student responds to the prompt with idea 18, stating that some of that energy gets trapped inside the car. In their final revision, the MI student prunes both of their initial mechanistic ideas and expresses idea 18, that the car is warmer inside because it is able to trap heat. Thus, from very different starting points, both the DI and MI students express the same, compelling mechanistic idea.

Progress of DI and MI students from pretest to posttest

On the two pretest/posttest items (Coal and Impacts), students demonstrated significant pretest (M=5.51, SD=1.09) to posttest (M=6.50, SD=1.35) learning gains while interacting with the unit ($t(69) = 6.39$, $p<0.001$, $d=0.764$).

Between pretest and posttest, on the *Coal* item, MI and DI students converged on explanations featuring mechanistic ideas about climate change. Each group experienced two interactions with the adaptive dialog that elicited their ideas about the mechanism of climate change. The mechanism students needed to explain the *Car* item used in the adaptive dialogs was the same as the mechanism students needed to explain in the *Coal* item. The DI students showed significantly greater gains (M=0.64, SD=0.91) than MI students (M=0.22, SD=0.95) on the *Coal* posttest item, which targets the climate mechanism ($t(68)=1.79$, $p=0.039$, $d=0.446$). Further, the posttest scores for the *Coal* item were not significantly different between DI and MI students, despite a significant

difference at pretest. This suggests that the dialog supported DI students to build upon their ideas to reach the same level of mechanistic explanation as their MI peers.

On the *Impacts* item, which prompted students to connect their understanding of the mechanism of climate change to explain why some groups experience more harmful impacts of climate change than others, there were similar pretest to posttest gains between the MI and DI students. The MI students had significantly higher posttest scores (M=3.49, SD=0.73) on the *Impacts* item than the DI students (M=2.8, SD=0.87, $t(68)=-3.55$, $p<0.001$). This result aligns with the lack of attention to policy and social factors in the adaptive dialog. The dialog did not elicit students' thinking about the issues in the *Impacts* item. It is possible that lack of attention to these ideas in the dialog, account for DI students' greater gains and higher average score on the *Coal* item than on the *Impacts* item.

Table 4
Representative examples of the second interaction with the adaptive dialog for students from DI and MI categories. Ideas detected by the model are italicized.

Time	DI Student	MI Student
Initial	It'll be warmer because the *sun is looking right inside the car.*	The *sun uses it's solar radiation to heat the Earth* by infared radiation *getting absorbed into the atmosphere and some reflects off the Earths atmosphere*.
Adaptive Guidance	Can you tell me more about how the Sun makes the inside of the car warmer than outside?	What happens to the energy from the sun when it is absorbed by the car?
Student Response 1	It will be warmer because the *car traps in all of the heat off of the sun*.	Some of the *energy is trapped inside the car* and some is released.
Non-adaptive Guidance	What's an idea you feel unsure about and chose not to include in your explanation?	What's an idea you feel unsure about and chose not to include in your explanation?
Student Response 2	I'm not unsure.	Nothing I'm confident about my answer.
Final Revision	The *car will be warmer* because the *heat from the sun will be trapped in the car making* it warmer than outside.	The *temperature inside the car will be hotter than the outside* because of the *car being able to trap heat* like the Earth does with the sun.

Conclusion and significance

The adaptive dialog shows promise as a strategy for eliciting students' ideas whether they start with descriptive or mechanistic ideas. It seems particularly useful to help students elaborate initial descriptive explanations. We found that students starting with descriptive ideas and starting with mechanistic ideas followed distinct paths during the dialog and achieved similar understanding of the mechanism of climate change by the posttest. The use of eliciting guidance was particularly impactful for DI students, who incorporated new mechanistic ideas into their initial explanations. This aligns with other research that suggests the importance of building on students' funds of knowledge as they construct understanding (diSessa, 2006; Linn, 2006; Hammer, 2000). The eliciting guidance supported MI students to articulate additional mechanistic ideas, which they often prioritized over their initial ideas during revision. Both groups increased their overall KI score at posttest.

The path taken by MI students suggests the potential of leveraging information about the KI score and the presence of multiple ideas when designing and assigning the adaptive guidance prompts. Students who started with one or more mechanistic ideas often articulated additional mechanistic ideas within the dialog. They pruned some of their initial mechanistic ideas to settle on promising final ideas. The KI Framework (Linn & Eylon, 2011) suggests that MI students need further support to distinguish among their mechanistic ideas and connect them into more complex explanations. Future iterations of the dialog could feature new guidance prompts that encourage students to sort out their mechanistic ideas and determine when they apply. This would enhance students' ability to draw connections among their mechanistic ideas to explain the *Car* phenomena.

Although the adaptive dialog was not aligned with the *Impacts* item, students made gains in Impacts KI scores, consistent with the overall focus of the unit. Both MI and DI students made progress in connecting their ideas about the causes of climate change to the social and political reasons that resulted in marginalized communities experiencing greater impacts of climate change. Further research is needed to determine whether an adaptive dialog using an idea detection model for the topics in the *Impacts* item would support greater student learning about the political and social factors that intersect with climate impacts.

Overall, the adaptive dialog was effective at drawing out students' ideas and enabled them to persist in reasoning about their explanations. The wide range of ideas expressed within the dialogs illustrates the value of the unit and dialog design. The dialogs reinforced the classroom culture established by our teacher participants and amplified their ability to both elicit students' descriptive ideas and respond to them.

References

Bang, M. & Medin, D. (2010). Cultural processes in science education: Supporting the navigation of multiple epistemologies. *Sci. Ed., 94*, 1008-1026.

Bradford, A., Gerard, L., Lim-Brietbart, J., Miller, J., Linn, M.C., (2022). Computational Thinking in Middle School Science. n C. Chinn, E. Tan, C. Chan, & Y. Kali (Eds.), Proceedings of the 16th International Conference of the Learning Sciences - ICLS 2022.

Carlone, H. B., Johnson, A., & Scott, C. M. (2015). Agency amidst formidable structures: How girls perform gender in science class. *Journal of Research in Science Teaching, 52*(4), 474-488.

diSessa, A. A. (2006). A History of Conceptual Change Research: Threads and Fault Lines. In R. K. Sawyer (Ed.), *The Cambridge handbook of: The learning sciences* (pp. 265–281). Cambridge University Press.

Furberg, A., & Silseth, K. (2021). Invoking student resources in whole-class conversations in science education: A sociocultural perspective. *Journal of the Learning Sciences*, 1-39.

Gerard, L., Bichler, S., Bradford, A., Linn, M. C., Steimel, K., & Riordan, B. (2022). Designing an Adaptive Dialogue to Promote Science Understanding. In C. Chinn, E. Tan, C. Chan, & Y. Kali (Eds.), Proceedings of the 16th International Conference of the Learning Sciences - ICLS 2022.

González, N., Moll, L. C., & Amanti, C. (Eds.). (2006). *Funds of knowledge: Theorizing practices in households, communities, and classrooms*. Routledge.

Hammer, D. (2000). Student resources for learning introductory physics. *American Journal of Physics*, 68(S1), S52–S59.

Linn, M. C. (2006). *The knowledge integration perspective on learning and instruction*. Cambridge University Press.

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

Liu, O. L., Lee, H. S., Hofstetter, C., & Linn, M. C. (2008). Assessing knowledge integration in science: Construct, measures, and evidence. *Educational Assessment*, *13*(1), 33-55.

Riordan, B., Cahill, A., Chen, J. K., Wiley, K., Bradford, A., Gerard, L., & Linn, M. C. (2020, February). Identifying NGSS-Aligned Ideas in Student Science Explanations. In Workshop on Artificial Intelligence for Education (AI4EDU@ AAAI).

Riordan, B., Jin, H., Lima, C., Steimel, K., & Yan, D. (in preparation). The application of automated scoring technology in learning progression assessment. In H. Jin, D. Yan, & J. Krajcik (Eds.), Handbook on Science Learning Progressions. Routledge.

Schulz, C., Meyer, C. M., Sailer, M., Kiesewetter, J., Bauer, E., Fischer, F., Fischer, M. R., & Gurevych, I. (2018). Challenges in the Automatic Analysis of Students' Diagnostic Reasoning. Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI).

Schulz, C., Meyer, C. M., Kiesewetter, J., Sailer, M., Bauer, E., Fischer, M. R., Fischer, F., & Gurevych, I. (2019). Analysis of Automatic Annotation Suggestions for Hard Discourse-Level Tasks in Expert Domains. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2761–2772.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practice, 31(1), 2–13.

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet?-A systematic literature review on chatbots in education. *Frontiers in artificial intelligence*, *4*.

Zacharia, Z. C., Manoli, C., Xenofontos, N., De Jong, T., Pedaste, M., van Riesen, S. A., et al. (2015). Identifying potential types of guidance for supporting student inquiry when using virtual and remote labs in science: A literature review. Educational Technology Research and Development, 63(2), 257–302.

Acknowledgements

Our work is funded by NSF grant 2101669: Using Natural Language Processing to Inform Science Instruction.