Assessing and Guiding Student Science Learning with Pedagogically Informed Natural Language Processing (NLP)

Authors: Marcia C. Linn, Libby Gerard
University of California, Berkeley

Abstract

Natural Language Processing (NLP) tools can score students' written explanations, opening new opportunities for science education. Optimally, these scores offer designers opportunities to align guidance with tested pedagogical frameworks and to investigate alternative ways to personalize instruction. We report on research, informed by the Knowledge Integration (KI) pedagogical framework, using online Authorable and Customizable Environments (ACEs), to promote deep understanding of complex scientific topics. We study how to personalize guidance to enable students to make productive revisions to written explanations as students conduct investigations with models, simulations, hands-on activities, and other materials. We describe how we iteratively refined our assessments and guidance to support students to revise their scientific explanations. We discuss how we explored hybrid models of personalized guidance that combine NLP scoring with opportunities for teachers to continue the conversation.

Funding

Funding for this research was provided by the National Science Foundation awards NLP-TIPS: Natural Language Processing Technologies to Inform Practice in Science (NSF project 2101669), STRIDES: Supporting Teachers in Responsive Instruction for Developing Expertise in Science (NSF project 1813713), and GRIDS: Graphing Research on Inquiry with Data in Science (NSF project DRL-1418423). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Introduction

Natural Language Processing (NLP) tools can score students' written explanations with pedagogically inspired rubrics (e.g. Kubsch et al., 2022; Linn et al., 2014; Zhai et al., 2020), expanding opportunities to scaffold student learning and investigate personalized guidance. We can score student written responses. The challenge is to make optimal use of the scores. We report on how we iteratively refined our instructional designs, informed by the Knowledge Integration (KI) pedagogy, to deepen student understanding of science. We describe how our recent investigations using NLP scores in units designed using an Authorable and Customizable Environment (ACE) can amplify teacher guidance and improve student ability to revise their scientific explanations.

Transforming Education with Technology

Starting in the 1970's, with the advent of the personal computer, there were widespread claims that technology would transform education (e.g. Darrach, 1970). However, most initial uses of technology in education mimicked the functionality of existing educational materials, often implementing a transmission model of education, rather than exploiting the affordances of emerging technologies (Linn, 2003). The first computers available in schools supported programming in BASIC or LOGO and offered designers limited resources for creating applications (Friedler et al., 1990; Mandinach, et al., 1986). In 1980 Seymour Papert laid out a perspective on constructivism and described the potential of teaching students to program in LOGO in Mindstorms (Papert, 1980). Many warned about the widening digital divide in educational opportunity as computers became available to students whose families could afford them (e.g., Linn, 2003; Lockheed & Frakt, 1984).

With increased access to more powerful computers for education, designers extended efforts to align innovations with pedagogical frameworks. One group created cognitive tutors primarily for mathematics and computer science guided by ACT* theory (Anderson, 1996), later incorporating the Knowledge, Learning and Instruction (KLI) pedagogy (Koedinger et al., 2012; Van Lehn, 2011). These tutors sought to achieve the same level of proficiency as typical instruction, in topics such as algebra, geometry, and LISP programming, while shortening the instructional time needed (Anderson et al., 1995). Empirical studies showed that students were learning skills in production-rule units and that the best tutorial interaction style was one in which the tutor provides immediate feedback, consisting of short and directed error messages. The cognitive tutors were effective in detecting and correcting student errors while students solved problems and they became commercial products (Koedinger & Aleven, 2016).

In science, many investigators built on constructivist pedagogy (e.g., Cognition and Technology Group, 1991; Gerard et al., 2015a; Inhelder & Piaget, 1987) with the goal of supporting inquiry learning. These groups designed technology-enhanced environments designed to promote self-directed learners who could plan a series of investigations (e.g., Lieberman & Linn, 1991).

In this paper we illustrate how the constructivist pedagogy, Knowledge integration, informed our use of NLP to design assessments and guidance. The KI pedagogy emerged from longitudinal and experimental studies of science and computer science instruction, designed to use

innovative technologies to improve student outcomes and teacher success (Clancy & Linn, 1999; Linn & Clancy, 1992; Linn, 1995; Linn & Hsi, 2000; Mokros & Tinker, 1987). It draws on research showing that students typically have multiple, incomplete, and fragmented ideas about scientific phenomena (diSessa, 2000; Smith et al., 1994). The KI pedagogy focuses on supporting students to analyze their own ideas, discover new insights into the phenomena, distinguish among these ideas, and reflect on their investigations (Linn & Eylon, 2011).

Authorable and Customizable Environments (ACEs) for Inquiry Science

Authorable and Customizable Environments (ACEs) that supported students to construct understanding emerged in the 1990s and enabled designers to scaffold science learning (Quintana et al., 2004). Rather than emulating typical textbooks, ACE design was informed by constructivist frameworks that emphasized inquiry learning (e.g., Linn & Eylon, 2011). ACEs were designed by partnerships of teachers, computer scientists, discipline experts, and learning scientists (e.g. Konings et al., 2014; Kyza & Agesilaou, 2022; Shear et al., 2004; Slotta & Linn, 2009). Partnerships included Concord Consortium (Concord.org, Molros & Tinker, 1987), Go-Lab (https://www.golabz.eu/, DeJong, et al., 2021), PhET (https://phet.colorado.edu/, Weiman et al, 2008), STOCHASMOS (Kyza et al., 2007) and the Web-based Interactive Science Environment (WISE, https://wise.berkeley.edu; Linn & Eylon, 2011). Many ACEs are free and open source, encouraging teachers and researchers to create experimental activities.

ACEs log student work and include scaffolds to guide students, making them potentially ideal for leveraging NLP tools. Further, ACEs take advantage of interactive visualizations of scientific phenomena including models or simulations and real time data collection, as well as creating ways to support hands-on investigations (Smetana & Bell, 2012). Studies of designs using inquiry to illustrate complex ideas in specific disciplines showed the value of using visualizations in Chemistry (Linn et al., in press) and other science disciplines (McElhaney et al., 2015). Many advocated combining virtual and hands-on activities to capitalize on the strengths of each format (de Jong et al., 2013). Many ACEs also feature collaborative tools (Ke & Hoadley, 2009; Matuk & Linn, 2018) and aspects of Learning Management Systems (LMS). ACEs can incorporate varied assessments including engineering designs (McBride et al., 2016), concept maps (Ryoo & Linn, 2012), and written explanations for complex questions (Tansomboon et al., 2017).

Researchers have been investigating ways to personalize guidance in ACEs taking advantage of logs of student interactions and responses to assessments (e.g. Gerard et al., 2015a; Puntambekar & Hübscher, 2005). Materials delivered by ACEs are often easy to use in experimental designs. They can support personalized instruction for each student as well as random assignment of students to conditions. Partnerships have built and tested multiple activities that are used by 1000s of teachers today. The major finding from a wide range of studies using ACEs is that inquiry learning is facilitated by personalized guidance and that

teachers benefit from tools to amplify their efforts to guide their students (e.g., Furtak et al., 2012).

Most ACEs are supported by systems such as WISE that are themselves open source and available on GitHub (https://github.com/WISE-Community). Materials delivered by ACEs are often easy to use in experimental designs. They can support personalized instruction for each student as well as random assignment of students to conditions. Embedded assessments enable teachers to monitor student progress during learning and to use student work when planning customizations (e.g., Wiley et al., in press).

As NLP tools have been incorporated into ACEs, the field has also begun to address issues of privacy, algorithmic bias, ethics, and equity (e.g. Higgs & Vahkil, 2019). For example, it is essential to develop NLP scoring algorithms with students who have the same cultural background as the students who will use the materials. The advances in ACEs support the possibility of transforming education with seamless school to home solutions.

The WISE ACE and KI Pedagogy

Reviews show the advantage of using pedagogies such as KI to guide the design of ACEs, assessments, guidance, and tools for teachers (Donnelly et al., 2014; Krajcik & Mun, 2014; Linn Donnelly, & Gerard, in press; Reiser et al., 2021). The WISE ACE offers tools that support designers to implement KI processes (Linn, Clark, & Slotta, 20). WISE *elicits student ideas* in multiple ways, often by posing dilemmas such as, “How do animals get energy from the sun?” Or by asking students to make predictions about a complex situation such as predicting the temperature within a car sitting in the sun on a snow-covered road. To *discover ideas* students often use models, virtual experiments, or hands-on investigations such as using a temperature probe to measure the temperature of objects in the room and comparing the measured temperature to how the object feels. To *distinguish ideas*, students might conduct virtual experiments by varying the amount of CO2 in the atmosphere, conduct hands-on experiments with temperature probes to compare the insulating properties of cups made of different materials, or use the Idea Basket to compare their explanations to those of their peers (e.g., Matuk & Linn, 2018). To *reflect on their ideas* students might write essays, make concept maps, or sort materials by some property. The WISE ACE logs all the students’ activities and can provide real-time personalized guidance based on student responses.

KI Assessments

KI assessments are embedded in WISE instruction, including as pretests and posttests. They may feature designing experiments using a virtual system (McBride et al, 2016), making a concept map (Ryoo & Linn, 2012), or writing explanations of complex situations where they link ideas with evidence (e.g.,Tansomboon et al., 2017; Vitale, Appleaum, & Linn, 2019). KI rubrics analyze student explanations for promising ideas, links between ideas justified by evidence, and multiple links between ideas (see example question and rubric in Table 1). Rather than rewarding only the right answer, KI rubrics reward students for sorting out their disparate ideas and using evidence to justify the ideas they incorporate into their explanations.

WISE units have always featured explanation items, due to their value in developing understanding. Generating explanations is more effective than answering multiple choice questions during learning (Richland et al., 2007). Research shows that generation items are also better predictors of long term retention than recall items, even for straightforward material (Bertsch et al., 2007). Further, a comparison of multiple choice and KI items covering the same material showed that KI items were better than multiple choice at discriminating between high and low scorers; they also captured nuances of progress more effectively (Lee et al., 2011).

Nevertheless, most science assessments use multiple choice questions that require recall of details and some form of problem solving (e.g. Third International Mathematics and Science Assessment, PISA) often informed by an information processing pedagogy (e.g., Anderson, 1996). This has led to extensive classroom practice on Multiple Choice factual questions that are often embedded in textbooks. It sends a message to teachers that recall is a major component of science learning. Yet, little of this information is retained as indicated by many assessments of adults. Further, investigations of public understanding of science consistently reveal weak and fragmented understanding of crucial science concepts (e.g., Weber & Stern, 2011).

Table 1: KI Rubric and Automated Guidance for "Cancer" explanation in a Mitosis unit.

***Item Prompt**:* Humans have a control mechanism that regulates cell division. When that control mechanism is broken, cells are allowed to divide out of control. As we have seen, this can lead to cancer. Now that we have successfully completed our investigation, let's use what we know to design a new drug to treat cancer. Which phase of mitosis would you have your drug target to stop cancer growth? [MC - NOT scored]. *Explain the effect your drug would have on the different parts of the cell in that phase, and how this would help keep cancer growth under control.*

Key Ideas

Describes a cell organelle or phase to be affected by the medicine
- Ways students might say this: "My drug would make the spindle fibers not be able to grow"

Explains how the medicine will disrupt the function of/action related to the organelle
- Function of OR action for an organelle (e.g.chromosomes carry genetic info (function) OR chromosomes are pulled apart (related action)
- Ways students might say this: "keeping cancer cells' chromosomes from being pulled into equal portions of each new cell."

Mentions need for medicine to stop cell division/cancer growth
- Controlling cancer cell growth; Stopping cell division; Stopping cells from splitting; Stopping X from making new cells
- Ways students might say this: "if you stop the chromosomes from dividing then there wouldn't be any new cell"

Side effects of drug treatment and their cause.
- Brings together ideas "stops cell division" + "but will stop good and bad cells (or cancerous and non-cancerous or normal and bad)"
- Ways students might say this: "cancer drugs target any cell that is dividing, which means cuts heal slower, hair may not grow back."

Score	Criteria	Student Examples	KI Guidance
1	**Off Task:** Writes text, does not answer question	IDK	Think about mitosis. What phase of mitosis will your drug target and WHY? Look at the phases in Step 2.5. Then, write a new description of your drug below.

2	**Irrelevant/Incorrect** Incorrect/nonnormative ideas Vague response	it would help the cancer cells stop growing because the cells would die.	Think about mitosis. What phase of mitosis will your drug target and WHY? Look at the phases in Step 2.5. Then, write a new description of your drug below.
3	**Partial Link** Any ONE of the three key ideas is correctly explained. OK to mention additional, incorrect ideas.	my drug would either freeze or burn the cell at their point where the cell multiplies	Good start - you are moving in the right direction. Now, add details about what function of the cell is important to stop and WHY. Watch the phases in Step 2.6 to gather ideas. Then, write a new description of your drug below.
4	**Full Link** Any TWO of the three key ideas are correctly explained and linked**.**	I would stop this phase because if you stop the spindle fibers, the chromosomes just float around and the cell can't reproduce.	Great work! Now, think about side effects - how will your drug affect the body? Check out Step 3.2 to gather some ideas. Then, write a more detailed description of your drug below.
5	**Complex Links** All THREE key ideas are correctly explained and linked.	This phase is the beginning of cell division. If you stop the cell from duplicating the chromosomes, the cell will not divide.	Nice thinking! Now, think about the rate of cell division in different parts of the body. Where in the body will your drug have the greatest impact? Check out Step 1.4 to gather some ideas. Then, write a more detailed description of your drug below.

Although teachers would like to score each students' work, this is difficult in middle school when teachers often have six science classes per day each with 30 students. Teachers report spending up to 10 minutes per student writing personalized guidance when teaching an online inquiry unit - or 5 hours per class (Gerard et al., 2015). Developing NLP to score student written responses initially involves a big commitment to collect and annotate the data, and build the model. Once developed, the NLP models can be used for many students as long as the new students resemble the students whose responses were used for training (Liu et al., 2014).

NLP Scoring

Our partnership uses c-raterML™, a tool developed by our collaborators at the Educational Testing Service (ETS) to score explanations for KI. The c-raterML system uses natural language processing methods to automatically score student written essays for KI. The system scores each student essay based on a 5-point knowledge integration rubric that rewards students for using evidence to make links among scientifically normative ideas. For example, as shown in Table 1, craterML assesses the degree to which students link ideas about cell organelle functions, the phases of mitosis, and health impacts from interrupting cell division, to explain how a drug they have designed will slow the spread of cancer.

craterML works by building a model of the linguistic features evident in human-scored student explanations at each knowledge integration score level. To build the human-scored data set, we collect over 1000+ student responses to the prompt from students in schools with demographics similar to those who will use the NLP-based guidance. This is done by working in sustained

partnership with teachers who are using the web-based curriculum units in which the explanation prompts are embedded. Two humans use the KI rubric to reach inter-rater reliability. They then score at least 1000 student written explanations to the prompt. c-raterML forms a statistical model based on its analysis of the given human scored data set. The c-raterML™ scoring has demonstrated satisfactory agreement with human scoring for constructed response items in inquiry science, meaning they demonstrated a sufficient quadratic-weighted kappa (K_{QW}) (Liu et al., 2014; Liu et al., 2016; Riordan et al., 2020). The quadratic-weighted kappa coefficient indicates the percentage of score agreement between the automated score and the human-assigned score, beyond what is expected by chance. It uses a range from -1 to 1 in which -1 indicates poorer than chance agreement, 0 indicates pure chance agreement, and 1 indicates perfect agreement (Fleiss & Cohen, 1973). Models that result in a coefficient above .75 in testing on a novel data set of student responses are deemed sufficient for use in instruction. This level of agreement is equivalent to rater agreement between typical trained humans.

The resulting NLP models can be deployed in the WISE units. They enabled our partnership to score KI explanation items that ask students to generate their ideas instead of using multiple choice items that often rely heavily on recall of information and knowledge of advanced vocabulary to differentiate among responses. Establishing KI scores is the first step. A big question is how best to use the scores.

KI essays and NLP tools have the potential to amplify the guidance of teachers by strengthening the guidance that ACEs can offer students (e.g., Gerard et al., 2015b). The WISE ACE has incorporated NLP and Machine Learning (McBride et al., 2016). In the next section we illustrate how our partnership has iteratively refined personalized guidance for science phenomena such as thermodynamics, photosynthesis, and plate tectonics.

Design Research: Personalized Guidance for Revision

We report on the results from the design research conducted by our partnership to find effective ways to use NLP scores to promote productive revision of scientific explanations. We conducted multiple iterations of the guidance across many WISE units on varied topics in middle school classrooms (see Table 2). We use mixed methods including quantitative methods to analyze both comparison studies and pre/post studies and qualitative methods in observational and interview studies.

Partnership

Our partnership had many participants who met regularly at school site meetings, professional development workshops, and on-campus seminars. Each partner contributed to the outcomes and respected the expertise of the others. We partnered with teachers and their students in over 12 participating middle schools. Partners included the computer scientists and software designers who created the WISE ACE and refined it to incorporate NLP scoring into a set of 12 units on topics including mitosis, photosynthesis, chemical reactions, thermodynamics, plate tectonics, and global climate change. Partners also included learning science researchers, psychometricians, and professional developers. We partnered with experts in NLP at ETS in

Princeton. The partners each conduct design research consisting of iterative refinement of guidance for revision with the goal of using NLP tools to personalize guidance for students and to help teachers guide their students to revise their ideas (e.g., Linn et al., in press; Wiley et al., in press). This work is both informed by the KI pedagogy and intended to strengthen the KI pedagogy. These studies have enabled us to refine our KI assessments, rubrics, instructional frameworks, professional learning tools, and the WISE ACE.

In partnership meetings we analyzed the results of each study and discussed ways to improve the personalized guidance, drawing on the KI pedagogy and related research. We focused on guiding students to revise their explanations based on research showing that generating explanations is a powerful way to build student understanding of their experimental investigations (e.g., Krist, 2020), and that productive revision can deepen student understanding (Berland et al., 2016; Hayes & Flower, 1986). This also aligns with the reported value of explanation in laboratory science as supported by ethnographic studies of scientific communities (e.g., Latour, 1987). We recognized that revision is difficult to motivate and often superficial (Crawford et al., 2008; Freedman et al., 2016).

Assessments of Revision

We used embedded explanations requiring knowledge integration scored by KI rubrics that reward students for linking ideas with evidence, reinforcing self-directed learning and knowledge building (e.g., Scardamalia & Bereiter, 2006). We scored the initial response using NLP and assigned guidance. We then scored the resulting revision in several ways. First, we noted whether the student revised their response. Then we scored the revised response using the KI rubric. In addition, in later studies we analyzed the nature of the revisions students made including noting whether students tacked on ideas, paraphrased their initial ideas, made grammatical improvements, or revised their reasoning.

In some studies we use a pretest/posttest KI revision item to assess student progress in revision across the unit. The KI revision items asked students to write an explanation, gave the student guidance, and asked the student to revise their response. Since most 5 to 12 day units had one NLP item, we did not expect the limited opportunity for revision during the unit to have a big impact on the KI Revision items between pretest and posttest.

Guidance Designs

We designed guidance to encourage students to build on their insights and observations in conjunction with new information, following the KI pedagogy. Students responding to KI guidance can use resources in the unit to find the evidence they need to determine which ideas are most useful and valid. As shown in Table. KI Rubric and Guidance, the guidance included a prompt with a link to visit relevant evidence within the unit. Rather than relying on authorities, the guidance encourages students to gain appreciation for relying on evidence to refine their knowledge. Ultimately KI guidance promotes cumulative understanding. Students who become better at KI become able to evaluate new arguments and see if they align with the evidence available to them.

In this paper we synthesize our design research studies to illustrate how taking advantage of emerging NLP technology and analyzing the impact of each revision is strengthening our understanding of personalized guidance. Each iteration of the instruction led to insights into the factors contributing to productive revision and to ways that personalized guidance can amplify the impact of teachers during inquiry instruction.

Initial Studies: Comparing Guidance Designs for KI

In our initial studies, we developed adaptive knowledge integration (KI) guidance for students' written arguments using craterML (Linn et al., 2014; See Table 1). The goal of the guidance was to prompt productive revision of student ideas. We designed the adaptive KI guidance to align with the KI pedagogy. It built on the current student answer and was intended to enable the student to move to the next level of the KI rubric. It included 4 parts: (1) Acknowledgement of the students' current ideas, (2) a question about the key missing or non-normative idea, (3) a suggestion to revisit related evidence in a dynamic visualization, and (4) a prompt asking the student to use the evidence they've gathered to generate an improved response (see Figure 1).

In our four initial studies, we found that the adaptive KI guidance was more effective in improving students' knowledge integration abilities, relative to other types of guidance typically used in middle school classrooms (Table 2, Rows A-D). We compared KI guidance to simulated teacher guidance (e.g. Redo. What does increased carbon dioxide do to global temperature?), generic guidance (e.g. Go back and review the visualization to improve your answer), and specific guidance (e.g. Light energy transformed into _____ kind of energy). Studies of the logged revisions and student navigation indicated that in comparison to the other forms of guidance, the KI guidance was more likely to support students to revisit specific evidence in the unit when revising, and to integrate a new idea into their initial response.

Table 2: Classroom Studies of NLP-based KI Guidance for Student Written Explanations. A-D Initial Studies in *italics*; E-H: Refinement of Guidance; I-J: Teacher Alerts; K-M: Modeling Revision with the Annotator.

Citation	Study Design & Topic	Impact on Item Revision	Impact on Pre/post Revision	Prior Knowledge Interaction
A: Tansomboon et al.. (2015). AERA	KI vs Simulated Teacher Guidance,1 round —- Global Climate Change (GCC)	KI more effective	No difference in KI gains between conditions. Students who integrated an idea (normative or non-normative) when revising during instruction made greater pre-post gains than students who made superficial revisions	No interaction detected

B: Gerard et al.. (2015b). Ed Psych Study 1	Teacher assigned KI v. Teacher assigned Generic, 1 round —- Mitosis, Chemical Reactions	KI more effective than generic guidance across contexts No effect for accuracy of NLP	No difference between conditions.	n/a
C: Gerard et al. (2015b). Ed Psych Study 2	KI Automated v. KI Teacher assigned, 1 round —- Cell Respiration	No difference in KI gains No effect for accuracy of NLP Took teachers 1-2 minutes to assign guidance for each student	No difference between conditions	n/a
D: Gerard et al. (2017). ESERA	KI v. Simulated Teacher guidance, 1 round —- Photosynthesis	No difference in KI gains between conditions. With KI guidance, more likely to integrate ideas when revising	Greater pre/post gains for low prior knowledge with KI guidance Low prior students who integrated an idea (normative or non-normative) when revising made greater pre-post gains than students who did not integrate when revising.	KI more effective for low prior on pre-post gains
E: Vitale et al. (2016).	KI v. **Specific guidance**, 2 rounds —- GCC	Slight advantage for specific guidance during instruction, not significant	Advantage of KI guidance for pre to post gains on essay item, and delayed posttest; Correlation between time spent revisiting visualization and pre/post gains	No interaction detected
F: Gerard & Linn. (2016b). AERA	KI Guidance - **Revision Rubric Categories**, 1 round —- Photosynthesis	Sig gains with KI guidance High prior students are more likely to integrate ideas when revising - low prior likely to add disconnected ideas.	Students who integrated ideas when revising made greater pre/post gains in School A. No diff in School B (73% in School B did not revise at all)	High prior students more likely to integrate ideas when revising.

G: Tansomboon et al. (2017). IJAIED - Study 1	Student name + **Transparent** KI v. Typical KI guidance, 2 rounds —- Thermodynamics	Transparent more effective. [No difference between students who engaged in 1 round versus 2 rounds of revision suggesting indication of progress did not impact outcomes.	No difference between conditions. Transparent more effective for low prior knowledge, sig higher scores at posttest	No interaction detected on embedded; Transparent more effective for low prior on pre/post
H: Tansomboon et al. (2017). IJAIED – Study 2	**KI Planning v. KI Revisiting** Guidance, 2 rounds —- Thermodynamics	No sig. gains in revision. Students in revisit more likely to revisit evidence; Students in planning more likely to make substantial revisions	No difference between conditions.	No interaction detected
I: Gerard & Linn (2016). JSTE	KI + Teacher **Alerts** v. KI, 2 rounds —- Photosynthesis	KI + Teacher alerts more effective for low prior knowledge in School A [no difference in School B]	Greater pre/post gains for KI + Teacher alerts for low prior knowledge in School A, than 2 rounds KI guidance [no difference in School B]	KI + Teacher Alerts more effective for low prior knowledge
J: Gerard et al. (2019). IJCSCL	Teacher adaptive KI + **teacher alerts** v. 2 rounds of KI guidance for low prior [Conditions did not hold in classroom study - only 2 pairs received a teacher alert] —- Plate Tectonics	Sig. revision gains. Teacher gave different guidance to low v. high, built on adaptive KI guidance. Teacher checked in with each group, high rate of revision	Sig pre/post gains	n/a
K: Gerard et al. (2016). ICLS	**Annotator** + KI KI, 2 rounds —- Photosynthesis	Annotator + KI made greater revision gains	Annotator + KI greater pre to post gains; Annotator + KI greater revision gains on posttest revision essay Students who made integrated revisions when revising, made greater pre to post test gains	High prior students more likely to integrate ideas when revising on KI revision item.
L: Gerard & Linn (2022). Computers & Education	KI 2 rounds v. **Annotator** + KI —-	Annotator + KI more effective on revision gains; Annotator + KI resulted in more integrated revisions	Annotator + KI greater revisions on posttest KI revision item	Annotator + KI more effective for low prior on

	Photosynthesis, Plate Tectonics			embedded and KI revision item
M: Linn & Gerard (in press) This Paper	**Annotator** + KI, 2 rounds	Annotate own explanation vs. Annotate fictitious student explanation	Both conditions revise, gain on KI revision item; Annotate fictitious student greater revisions	Fictitious student condition created more unique labels.

Refinement of Guidance

The initial studies established that our KI guidance was as effective as guidance from experienced teachers and more effective than typical guidance or completion guidance. In our refinement studies we sought to improve on the initial designs (Table 2, rows E-H).

Specific Guidance. We explored the role of KI versus Specific guidance where students were told the right answer (Vitale et al., 2016). Consistent with other research on learning, we found that specific guidance was as effective as KI guidance during instruction and that KI guidance was more effective than specific guidance for promoting durable understanding as measured by a delayed posttest (Richland et al., 2007).

Types of Revisions. We analyzed the types of revisions students made and found that they were often superficial. Students added ideas rather than thoroughly integrating the new information (Tansomboon et al., 2017). These findings are consistent with related research on writing and revision. When given feedback from teachers or peers, using technology tools such as collaborative Google Docs, students most often make minimal or superfluous changes to their science explanations (Freedman et al., 2016; Sun et al., 2016; Zheng et al., 2015; Zhu et al., 2020). Learners tend to make changes to spelling and grammar rather than to revise for meaning (Bridwell, 1980; Fitzgerald, 1987; Strobl et al., 2019; Zhu et al., 2020).

Transparency about Guidance. Some students did not recognize that the guidance was personalized and dismissed it (Tansomboon et al., 2017). Students' uncertainty about whether the guidance was personalized is consistent with beliefs about computers when the study was conducted. To help students appreciate that the guidance was personalized to their ideas we made the NLP process more transparent. We added student names to the guidance. We explained how the computer read their response, compared their response to the responses from 1000s of other students of the same grade level, and then selected guidance to address their distinct science ideas. We found that the transparent guidance condition led to greater rates of revision particularly for students who initially displayed low prior knowledge (Tansomboon et al., 2017). This extended prior research suggesting that when students are challenged, they are more likely to engage and persist if they perceive the guidance they receive as connected to their reasoning (Shute, 2008). We altered the KI guidance interface to always provide transparent guidance.

Reflection on Refinements. Analyzing the overall effectiveness of guidance in these refinement studies, we noted that although the KI guidance helped many students to integrate new evidence into their explanations and strengthen the links among their ideas, there were

limitations. Many students still struggled to use the guidance to revise their arguments--only about half of the students were able to make productive revisions (Gerard et al., 2016). In one study, over 50% of students who received automated guidance either did not revise their answers or only made surface-level changes without adding a new idea (Tansomboon et al. 2015). In addition, integrating new ideas when revising was most challenging for students who initially displayed low prior knowledge (Gerard & Linn, 2016b). This resonates with prior research findings that when confronted with contrasting evidence, students tend to ignore the evidence and restate their own perspective, consistent with confirmation bias (Clark & Chase, 1972; Höttecke & Allchin, 2020). Further, in student interviews conducted during guidance studies, some students reported that they preferred their teacher's guidance over the automated KI guidance because their teacher gave feedback that was specific to their response.

Combining Teacher and Personalized Guidance

To address the challenges faced by low prior knowledge students we tried alerting teachers to guide students who were stuck (Table 2, rows I-J). The partners decided on the conditions under which they wanted alerts. Typically, teachers wanted alerts when students made two attempts at revision without any progress or continued to express vague ideas (level 2 on KI rubric). We designed alerts which showed up on the students' computer screen. Teachers could see the alert as they circled the classroom. Students could keep working while the alert showed on their screen, and the teacher could come to talk with the student about the item (Gerard & Linn, 2016a). We found that the alerts led to gains in one school and not in the other school, suggesting that the process needed fine tuning.

Analyzing Revision Strategies and Modeling Revision

Our initial analysis of the nature of student revisions suggested the need for deeper understanding of how students were envisioning revision. We systematically investigated how students were revising their science writing based on the KI guidance, and what kinds of revisions to science arguments led to building coherent, long term science understanding (see Figure 1). To characterize how students revised their science writing based on the KI guidance, we analyzed students' writing in their initial and revised explanations and identified what changes, if any, students made to their writing (e.g. Gerard et al., 2016; Tansomboon et al., 2017).

In this coding process we noticed qualitatively different patterns in the kinds of revisions students were making after they received KI guidance. Specifically, some students made integrated revisions while others tacked on ideas or did not revise at all. Those who integrated ideas when revising during instruction, were also making greater pre to post test gains (Gerard, & Linn, 2016a; Tansomboon et al., 2015). We developed an emergent coding scheme that captured the patterns we observed: those who integrated new ideas when revising their writing, those who integrated redundant ideas or paraphrased what they had said initially, those who added new but discrete ideas, and those who made no changes at all (see Figure 1; Gerard & Linn, 2022). In coding students' writing revision strategies, we evaluated only the changes in the students' science writing, not the scientific accuracy of the change. Consistent with the KI

pedagogy, we hypothesized that making connections among ideas would be a more productive learning strategy than accumulating more discrete ideas or not refining the ideas at all.

We found that the type of revision strategy impacted learning outcomes (Tansomboon et al., 2017; Gerard et al., 2016). For example, in one study students wrote a short essay in a photosynthesis unit and received one round of KI guidance. We coded students' initial and final (after receiving the guidance) short essays in the unit, and student responses on pre/post test short essay items using knowledge integration rubrics. We found that students who integrated ideas when revising their essay during instruction made greater pre to posttest gains on the short answer items than those students who added ideas when revising on the essay activity during instruction, or those who chose to make no changes at all. The difference in pre/post test gains between those who integrated ideas and those who did not, was significant on the Energy Story pre/post item (Integrated n=181, M=.81, SD=1.16; Did Not Integrate n=159, M=.43, SD=.96; t(338)=3.19, p=.002). These results suggested that the students who made no attempt to integrate ideas lacked a model of the revision process.

Figure 1: Type of Revision Response to KI Guidance (Gerard & Linn 2022).

Description	Initial Response	Revised Response [***bold italics is revised idea***]
No Revision		
Integrated New Connects new idea(s) to initial ideas. The new idea builds on what was stated in initial response, by elaborating, extending, or contrasting.	Heat goes up into the atmosphere then cools down then goes back down into earths core and repeats the process.	Heat goes up into the atmosphere ***where the density is higher*** then cools down then goes back down into earths core ***where it is less dense*** and repeats the process.
	The heat makes it less dense, as it flows away from the heat source it becomes less dense. It comes back down to the heat source after being in circulation.	The heat makes it less dense, as it flows away from the heat source it becomes less dense. It comes back down to the heat source after being in circulation. ***The reason it goes from top to bottom is because it looses its density as it goes to the bottom. When it looses its density it gets lighter and floats to the top. When it gets heavier it sinks to the bottom. The process is ongoing.***
Integrated Redundant Adds an idea that repeats initial idea or paraphrases; does not add new science idea.	The bottom of the lamp is hot like the core. The blob is like the convection currents. At the bottom the blob is heated becoming less dense floating to the top. At the top it becomes less dense and goes back down.	The bottom of the lamp is hot like the core. The blob is like the convection currents. At the bottom the blob is heated becoming less dense floating to the top. At the top it becomes less dense and goes back down. ***Like the convection current in goes up and goes back down***.
	I think a lava lamp works by the heat in the lamp causing the blobs to go up word and then it gets more dense and then when the blobs go down it gets less dense.	I think a lava lamp works by the heat in the lamp causing the blobs to go up word and then the ***density increases*** and then when the blobs go down it gets less dense.

Disconnected New Writes entirely new response. Or, adds new idea with no edit to initial response, that does not connect to initial idea(s).	The heat from the lava lamp makes the blobs less and causes it to move easily. It's similar bc in earth's mantle it slowly comes out like a lava lamp	***When it's too hot at the bottom, it goes up, gets too cold, and goes back down, like a cycle.***
	We think that the blob of colored fluid goes up because of heat and density. Heat makes density less dense and density is what brings up the fluid.	***We think that, the blob of colored fluid goes up because of heat and density. Heat makes density less dense and density is what brings up the fluid. There is also more density on the top with low heat and less density on the bottom and high heat on the bottom***.

The Annotator. The challenge that integrating ideas posed for students suggested that some students were not sure what revision looks like. We designed the Annotator (See Figure 2) to provide students with an interactive model of integrated argument revision. The Annotator asks students to help a fictitious student make decisions about revision by placing premade labels on the students' response. The student also has the opportunity to author their own labels to guide the revision. An initial study of the Annotator showed that combining one round of adaptive KI guidance with one round of the Annotator was more effective in promoting integrated revision, especially for students who initially expressed "unintegrated ideas" and hence had received low KI scores (1 to 2), than providing multiple rounds of adaptive KI guidance (Gerard & Linn, 2022).

These findings documented the importance of providing a model of the revision process, especially for low prior knowledge students. When learners had the opportunity to select and place labels on the response of another student, they were more likely to revise their own response. Indeed, students often remarked that they were using the same strategy they used to choose a label when revising their own explanation. For example, one student reflected on their use of the Annotator in the Plate Tectonics unit: *"This way [the Annotator] gets your brain on what you need, like what she [fictional peer in Annotator] does not have…Placing the labels was useful [to revising in the next step] bc it had many things i didn't think about.*" Another student reflected on their use of the Annotator in the Photosynthesis unit: *"I realized I needed to expand more what I wrote."* Another student expressed: *"It helped set up a structure for my writing. I went back to our writing and thought about those questions."* Across the student interviews, across unit contexts, students reported how their experience using the Annotator helped them to notice gaps in their explanation, or to recognize a new idea they held to strengthen the links in their explanation.

An Experiment: Peer versus Self Annotator

To better understand the mechanisms underlying the benefit of the Annotator and to continue to refine the Annotator with a focus on fostering self-directed learning in revision we designed a version where students **annotated their own response**. We hypothesized that placing the pre-authored labels onto the explanation was the central mechanism promoting integrated revision. It (a) modeled for the student the process of distinguishing which key ideas in an explanation are missing by evaluating the response using the ideas in the labels and (b)

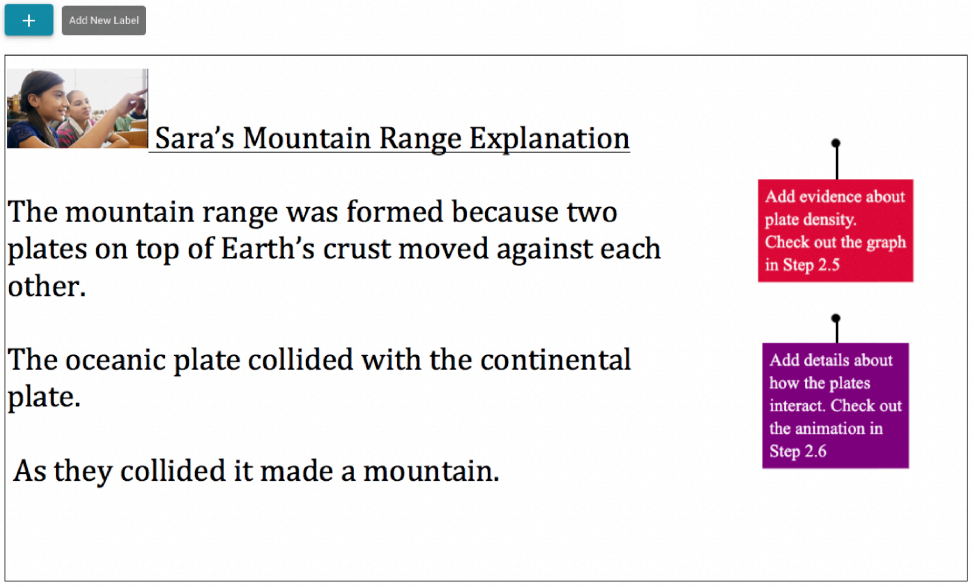
modeled how to link new ideas with existing ideas by determining where to place the labels onto the written response. It appeared that students could then apply this approach to their own explanation when using the KI guidance. To test this idea, we designed an iteration of the Annotator to support the student to annotate their own explanation - rather than a fictional peer's - with pre-authored labels. We conjectured that this may increase students' sense of autonomy and hence self-directed learning in the revision process while also promoting integrated revision.

We studied the impact of the self- and peer-annotations in an unpublished study. Students were asked to place pre-written labels on sections of an explanation to suggest areas for change or improvement. They were also given the opportunity to make self-constructed labels. The pre-written labels were designed to elicit evidence central to explaining the phenomenon that is most often missing in student explanations.

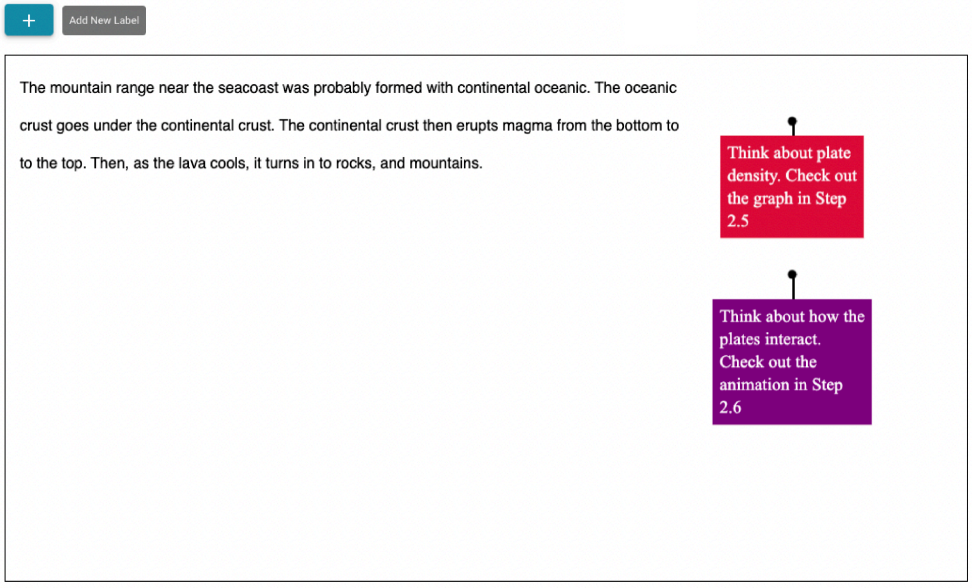
For example, an explanation prompt embedded in a unit on Plate Tectonics asks students to explain how Mt. Hood was formed (given a photograph of Mt. Hood on the Pacific coast). A pre-written label in the Annotator for a fictional peer's response to the Mt.Hood explanation, says "Think about plate density. Check out the graph in Step 2.5", since many students leave out this idea and it is central to understanding how the plates interact. Selecting the relevant labels and placing them in the written response encourages distinguishing of ideas in the response and in the labels, and the integration of new and prior knowledge, rather than novice practices of tacking on disconnected information. We compared this new version of the Annotator intended to strengthen student agency in revision to the initial Annotator design involving peer annotation. We hypothesized that instantiating the student's own essay in the Annotator would encourage students to view their essay as a scientific product and attend more carefully to each expressed idea, the connections among them, and possible gaps. Flower and Hayes (1980) showed that when students succeeded in analyzing the structure and argument of their essay, they were capable of making valuable revisions to their reasoning. (See Figure 2.)

Figure 2: Annotator tool to support revision. Students move pre-written labels to suggest ways an explanation can be improved. Students can also create their own labels.

Students Annotate Fictious Peer Explanation



Students Annotate Their Own Explanation



Methods. 5 teachers from 3 schools and their 678 7th-grade students participated. All students used the WISE Plate Tectonics unit (wise.berkeley.edu). For two activities embedded in the Plate Tectonics unit, all students wrote an explanation. Each prompt called for students to connect ideas about plate boundary interactions and convection to explain volcano formation. After writing their argument, when students moved to the next step, they were randomly assigned using the WISE branching technology to one of two conditions: (a) annotate their own argument or (b) annotate a peers' argument.

In the *Peer-annotator* version, an explanation by a fictitious peer named Sara was pre-loaded in the Annotator (Table 3). A peer explanation was selected at a KI level 3, to reflect a common student idea and missing evidence, making it generative for critique. The labels were pre-designed to be personalized to Sara's explanation, meaning that they elicited evidence which would link to an existing idea in the explanation. The pre-designed labels asked: (a) Add evidence about plate density. Check out the graph in Step 2.5, and (b) Add details about how the plates interact. Check out the animation in Step 2.6. The instructions also encouraged students to write their own label if they have another comment.

In the *Self-annotator* version, the student's written explanation was automatically imported into the Annotator (Table 4). The same pre-designed labels as in the 'Peer-Annotator' version appeared to the right of their explanation. While the labels were personalized to Sara's explanation, the labels raised key concepts that were general enough that we hypothesized one or both of the labels could likely be applied to improve most student written explanations.

In both conditions, students used labels to address gaps or inaccuracies in the explanation; revised their own work; and then had one opportunity to receive personalized KI guidance for their explanation and revise again. Students completed a pretest and a posttest before and after the unit with an item that called for students to write and revise. All students' written arguments were scored using 5-point knowledge integration rubrics that reward students for scientifically accurate links among ideas.

Findings. *Use of the Annotator to Critique Arguments.* We analyzed the students' annotations from one teacher in each of the three schools. Students were better able to identify and remedy gaps with scientifically accurate suggestions when annotating a fictional peer's explanation than when annotating their own. Student annotations in the peer-annotator condition were scored significantly higher, than those in self-annotator condition [Peer-annotator, M=2.03, SD=1.03; Self-annotator, M=1.33, SD=1.15; $t(282)=5.47$, $p<.0001$].

Students reported that the peer-annotator enabled them to gather new ideas. As one student stated: *"I like revising the classmates and ours was hard to revise because we're the ones who made them"*. Another student commented, *"I reviewed Sara's and so then I just added a sentence [to mine] because it gave me more information and then I put that into my own words"*. As seen in Table 3, the student added a new idea to their explanation about plate density after prompting the fictional peer Sara to consider this same idea. Teachers echoed the student perspective, noting that students were more likely to critique a peer's explanation than their own as they presume their own response is correct, and particularly for students with initially vague ideas, they may also be uncertain of what criteria to use to evaluate their own explanation.

In both Annotator conditions, students were given two pre-authored labels to use in annotating the argument, and they were also instructed to create their own new label if they identified an additional gap in the argument. Students in the peer-annotator condition were significantly more likely to create new labels during annotation (33% of students), than students in self-annotator condition (18%) [$X^2(1)=7.36$, $p=.007$]. For example, in Table 3, the student created a new label *"add evidence about how the plates move differently from in 2.1"*, prompting the fictional peer to consider how oceanic and continental plates interact differently than two continental plates

colliding. We coded the type of label students generated as either (a) general (e.g. add more), (b) add a new idea (e.g. what about convection currents?), or (c) fill a gap to clarify the mechanism (e.g. why does heat cause molecules to become less dense?). In both conditions, students primarily generated labels focused on filling a gap to clarify the mechanism in the explanation [peer-annotator, 72% of labels; self-Annotator, 68%]. Taken together, the analyses suggest that annotating a fictional peer's explanation may lead to greater student engagement in evaluating the ideas in a scientific argument, and in generating mechanistic ideas to strengthen the argument.

Table 3: Peer-Annotate Condition, Example of Student Writing, Annotation and Revision.

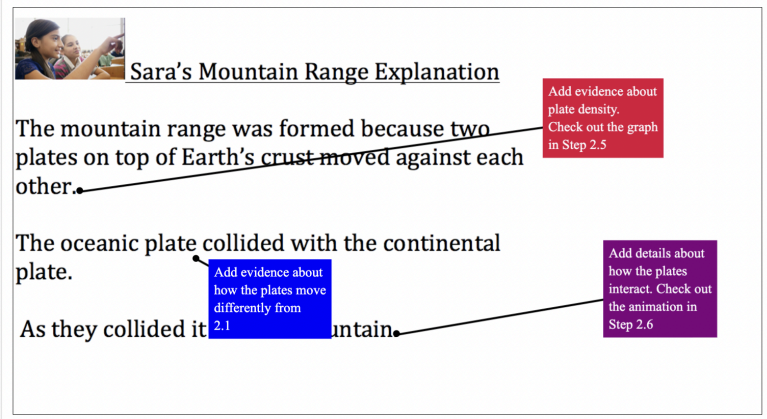
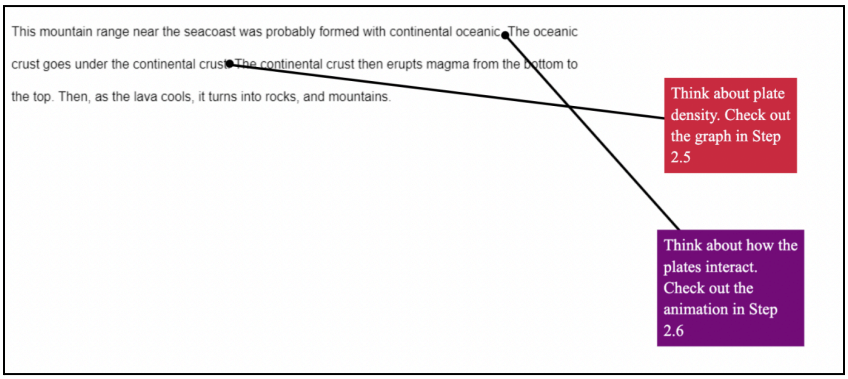
Event	Student Work
Initial explanation	**Initial:** The mountain range was probably formed by the Oceanic crust and Continental crust push against sediment that goes up. That sediment then turns into a mountain.
They place the labels where Sara should add the suggested ideas. They also add their own label encouraging the student to distinguish how this type of plate interaction, is different from the interactions that occur at a transform boundary and divergent boundary, as explored in Step 2.1	
After annotating Sarah's explanation, they revised their explanation. They added a new idea to their explanation that they had previously recognized was missing in Sara's explanation, based on their placement of the labels.	Revised 1: The mountain range was probably formed by the Oceanic crust and Continental crust push against sediment that goes up. **They push becuase one is more dense then the other,one goes under the continental crust**. That sediment then turns into a mountain.
The student then received **adaptive KI guidance** for their explanation.	**Adaptive KI Guidance**: *Elliott, add details to your explanation. How does the density of the two plates affect their movement? Check out for a hint. Then, expand your explanation.*
They further revise their explanation, clarifying the plate interactions due to their differing densities.	**Final explanation:** The mountain range was probably formed by the Oceanic crust and Continental crust push against sediment that goes up. They push because one is more dense ,**the oceanic crust goes under the continental crust,pushing the continental crust up. The crust then forms a mountain.**

Table 4: Self-Annotate Condition, Example of Student Writing, Annotation and Revision.

Event	Student Work
Writes explanation mid-way through the Plate Tectonics unit	**Initial Explanation:** This mountain range near the seacoast was probably formed with continental oceanic. The oceanic crust goes under the continental crust. The continental crust then erupts magma from the bottom to the top. Then, as the lava cools, it turns into rocks, and mountains.
The student's explanation is imported into the Annotator, with pre-authored labels on the side. In this case the labels hence are well aligned to gaps in the student's explanation. They place the labels in their explanation to indicate where to make a link to evidence.	
After the student annotated their own explanation, they incorporated a new idea to strengthen the link between plate density and subduction.	**Revised explanation after using the annotator**. "This mountain range near the seacoast was probably formed with continental oceanic. **The crust push into each other.** The oceanic crust goes under the continental crust**. This is because the oceanic crust is denser than the continental crust, so the oceanic crust subducts under the continental crust.** The continental crust then erupts magma from the bottom to the top. Then, as the lava cools, it turns into rocks, and mountains."
The student then received automated KI Guidance level 5	*Sam, nice thinking! Look over your explanation to be sure it addresses the density of the plates and how they interact. Revise your explanation as much as you think is needed.*
The student continued to clarify the role of density in plate movement.	**Final explanation**. This mountain range near the seacoast was probably formed with continental oceanic. The crust push into each other, **causing** the oceanic crust goes under the continental crust. **The oceanic crust is denser than the continental crust**. **Denser things sink, so that's why the oceanic crust went under the continental crust.** The continental crust then erupts magma from the bottom to the top. Then, as the lava cools, it turns into rocks, and mountains.

Embedded Revision Gains. Students used the guidance to significantly improve their explanations in the revision activity in both guidance conditions [AnnotatePeer Gain, M=.40, SD=.67, t(216)=8.76, p<.0001; AnnotateOwn, M=.37, SD=.68, t(218)=7.95, p<.0001]. Students made significant revision gains after using the Annotator in the first round of revision, and smaller gains in the second round of revision after receiving the KI guidance [AnnotatePeer (1st round) M=.27, SD=.65; (2nd) M=.13, SD=.61; AnnotateOwn, (1st) M=.30, SD=.65; (2nd) M=.07 SD=.57]. There was no main effect for the condition, suggesting that both critiquing a peer and one's own argument can strengthen student explanation writing and revising.

Pre to Post Test Gains. Students in both conditions made significant pre to post test gains [Gains: AnnotatePeer M=.56, SD=.88; AnnotateOwn M=.47, SD=.80] with no main effect for condition.

Discussion. Revising explanations is central to the iterative process of knowledge building in science yet it is unfamiliar and challenging to most learners (Berland et al., 2016; Mercier & Sperber, 2011). We created the **peer-annotator** to model the process of revising. We designed labels personalized to the response and also enabled students to write their own labels. We created the **self-annotator** to directly allow students to annotate their own response by using the labels or writing their own.

Our findings suggest that students benefit from a model of revision that helps them discern key science practice such as distinguishing criteria to critique a scientific argument. Students wrote more labels in the peer than the self condition, suggesting that the personalized labels in the peer annotator modeled revision. The peer annotator was more effective than the self condition for promoting revision of students' initial response. Students reported that they were more likely to gather new ideas that they could then apply to their own explanation.

This also suggests that students in the self-annotate condition did not see the pre-authored labels as helping them to identify new ideas to incorporate into their response. An important difference between the conditions was the design of personalized pre-authored labels for the fictional peer's explanation. The labels for the self-annotation condition were the same as in the peer-annotator condition and hence not personalized to the student's response.

This raises a question about the design of the Annotator. One question concerns the labels. Would students benefit from labels personalized to the response they are annotating? Can we use NLP to design personalized labels for students' own responses? Would using an NLP model to identify pre-authored labels for the students' own explanation enhance the self-annotator condition? Personalizing the labels may support students to (a) see how the model of revision is aligned to their ideas and hence elicit greater engagement in the revision process - building on Tansomboon et al, (2017), and (b) help the student distinguish between their ideas expressed in their explanation and those suggested by the personalized labels to determine what evidence to pursue fill a gap or clarify a link, to create a more coherent explanation.

A second question concerns the limitation of the model of revision in the Annotator. Do students need a model of the metacognitive processes of considering alternatives for revision? How could we design an annotator or another tool that enables students to diagnose that they need to distinguish among their ideas rather than tack another idea on to the explanation? Perhaps we can design a Metacognitive Annotator to engage students in distinguishing among possible revisions, some that tack on ideas and others that integrate evidence.

Discussion: Using NLP to Improve Personalized Guidance

Advances in NLP offer designers new opportunities to improve instruction. In our work, we were guided by the KI pedagogy as well as the insights of expert teachers to test and refine personalized guidance for KI items. We initially focused on moving students to the next level of the KI rubric. In our initial studies, we were able to approximate the guidance of expert teachers. And, like expert teachers, the personalized automated guidance had impacts and limitations.

The guidance significantly improved responses to the KI items. KI guidance was more effective than specific guidance, leading to durable understanding as measured by a delayed posttest.

Some of the limitations resulted from the automated nature of the guidance. Students prefer the guidance of their teachers, often saying that their teachers were more likely to provide the right answer. They suspected the automated guidance was not personalized. By being transparent about how automated guidance was designed, we were able to reduce distrust and increase the impact of automated guidance.

Some of the limitations aligned with prior research. Revision is difficult and superficial revisions are common. Our guidance had these same limitations. To make progress, we combined automated guidance with alerts to teachers about students who were struggling. This process has promise. We have initiated a line of work involving teacher dashboards that responds to teacher interest in more nuanced information about their students than simply an alert. They would like information about the whole class as well as about the needs of individual students so they can target their guidance to the needs of each student (Wiley et al., in press).

We also did a detailed analysis of the types of revisions students made and identified additional opportunities to improve personalized guidance. Some students did not have a clear understanding of the nature of revision. We designed the Annotator to model revision. The Annotator was helpful, especially for students who started with low prior knowledge and therefore were likely to lack a model of revision. To explore ways to improve the Annotator, we compared the situation where students annotated the response of a peer to a condition where they annotated their own explanation. We found that students, as anticipated, had difficulty annotating their own responses, consistent with work on the limitations of metacognition. An important difference between the conditions was that students placed a personalized pre-authored label to annotate the peer explanation. We identified some directions for future work on a self-annotator. For example, we hypothesize that an NLP model to identify pre-authored labels for the students' own explanation would enhance the self-annotator condition. We will also explore ways to design a MetaCognitive Annotator.

Design Research: Reflection

Emergent technologies offer educators opportunities to improve instruction. Finding optimal uses of these technologies often takes many design iterations. The iterations are informed by deliberations of the design research partners; analysis of the logged data that provides detailed insights into the interaction between student thinking, guidance, and revision; and reflection on the way results align with the underlying pedagogy. This paper reports on ways that NLP tools have been refined to improve student learning. The results show that personalized guidance aligned with the KI pedagogy emphasizing rewarding students to integrate their ideas has advantages for long term retention as well as the development of self-directed learners. This aligns with other work on self-directed learning (e.g., Scardamalia & Bereiter, 2006).

This design research illustrates how partnership refinement of guidance can improve student learning. By combining the expertise of each partner we were able to gain insight into ways to promote revision. We were inspired by the excellent guidance provided by expert teachers. We

benefited from the insights of psychometricians to design KI items that require students to generate arguments and that measure how students respond to personalized guidance with KI revision items. We refined the adaptive guidance that both promotes revision and encourages self-directed exploration of scientific evidence by conducting whole partnership reflections at professional development workshops (e.g. Gerard et al., 2022b). We were able to realize often nascent ideas when the software designers brought prototypes of the Annotator, new designs for discussion tools, and refinements to the interactive models to partnership meetings. We worked closely with the NLP designers to clarify the strengths and limitations of early models and improve accuracy (e.g., Riordan et al., 2020).

This design research tested forms of personalized guidance informed by KI pedagogy and strengthened our understanding of KI pedagogy as a result. By analyzing the ways that students respond to requests to integrate their ideas, this research revealed opportunities for refining KI design recommendations. We found that students often had no experience evaluating and revising science explanations, emphasizing the need for engaging students in finding gaps in their arguments. This resonates with earlier work on metacognitive reasoning and KI (e.g. Linn et al., 2004). We found that students often chose to tack on an idea rather than to distinguish it from their other ideas. This finding resonates with other studies of KI that revealed the need for more emphasis on distinguishing ideas (e.g. Gerard et al., 2020; Ryoo & Linn, 2012; Vitale et al., 2019).

NLP technology has the potential to amplify the role of the teacher by providing automated guidance to students and identifying the learners who would most benefit from teacher guidance. Building on the ways that successful teachers guide students, we show how NLP tools implemented in ACEs can strengthen science instruction. Initially, we diagnosed student performance using a KI rubric and designed guidance intended to enable the student to revise their explanation and achieve the next level of the rubric. This was helpful but many students floundered or did not revise at all. Based on observations of how teachers interact with individual students, we can envision the potential of hybrid models of personalized guidance that combine NLP scoring with opportunities for teachers to continue the conversation. Guidance embedded in the unit can encourage the student to strengthen the links between different pieces of evidence to explain a phenomenon. The guidance may also serve as a conversation starter between teacher and student or student and peer thus combining automated and human guidance. Building on the findings from studies combining adaptive KI guidance with a teacher alert, that alerts the teacher in real-time to students whose explanation was scored by the NLP below a set threshold, hybrid models optimize referrals to peers, teachers, or an alternative approach such as a computer-student dialogue. Our current work takes advantage of new NLP models designed to identify specific ideas rather than KI levels and could support these types of dialogues (Gerard et al., 2022a).

This design research program illustrates the process of iterative design and the ways it has benefitted student learning. Many challenges remain. As we noted initially, NLP can provide scores for student work, the challenge is figuring out what to do with the scores.

References [2724 words]

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist, 51*(4), 355–365. https://doi.org/10.1037/0003-066X.51.4.355

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive Tutors: Lessons Learned. *Journal of the Learning Sciences, 4*(2), 167–207. https://doi.org/10.1207/s15327809jls0402_2

Azevedo, R. (2005). Computers environments as metacognitive tools for enhancing learning. *Educational Psychologist, 40*(4), 193–197.

Berland, L. K., Schwarz, C. V., Krist, C., Kenyon, L., Lo, A. S., & Reiser, B. J. (2016). Epistemologies in practice: Making scientific practices meaningful for students. *Journal of Research in Science Teaching, 53*(7), 1082–1112. https://doi.org/10.1002/tea.21257

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*(2), 201–210. https://doi.org/10.3758/BF03193441

Clancy, M. J., & Linn, M. C. (1999). Patterns and pedagogy. *SIGCSE Bulletin, 31*(1), 37–42.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*(3), 472–517. https://doi.org/10.1016/0010-0285(72)90019-9

Cognition and Technology Group. (1991). Technology and the Design of Generative Learning Environments. *Educational Technology*, *31*(5), 34–40.

Crawford, L., Lloyd, S., & Knoth, K. (2008). Analysis of Student Revisions on a State Writing Test. *Assessment for Effective Intervention, 33*(2), 108–119. https://doi.org/10.1177/1534508407311403

Darrach, B. (1970, November 20). Meet Shaky, the first electronic person. *LIFE Magazine, 69*(21), 58B-68B.

de Jong, T., Gillet, D., Rodríguez-Triana, M. J., Hovardas, T., Dikke, D., Doran, R., Dziabenko, O., Koslowsky, J., Korventausta, M., Law, E., Pedaste, M., Tasiopoulou, E., Vidal, G., & Zacharia, Z. C. (2021). Understanding teacher design practices for digital inquiry–based science learning: The case of Go-Lab. *Educational Technology Research and Development, 69*(2), 417–444. https://doi.org/10.1007/s11423-020-09904-z

de Jong, T., Linn, M. C., & Zacharia, Z. C. (2013). Physical and Virtual Laboratories in Science and Engineering Education. *Science, 340*(6130), 305–308. https://doi.org/10.1126/science.1230579

diSessa, A. A. (2000). *Changing minds: Computers, learning and literacy*. MIT Press.

Donnelly, D. F., Linn, M. C., & Ludvigsen, S. (2014). Impacts and Characteristics of Computer-Based Science Inquiry Learning Environments for Precollege Students. *Review of Educational Research, 84*(4), 572–608. https://doi.org/10.3102/0034654314546954

Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction, 19*(3), 243–266. https://doi.org/10.1007/s11257-009-9063-7

Fleiss, J. L., & Cohen, J. (1973). The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educational and Psychological Measurement*, *33*(3), 613–619. https://doi.org/10.1177/001316447303300309

Freedman, S., Hull, G., Higgs, J., & Booten, K. (2016). Teaching Writing in a Digital and Global Age: Toward Access, Learning, and Development for All. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of Research on Teaching* (5th ed., pp. 1389–1449). American Educational Research Association.

Friedler, Y., Nachmias, R., & Linn, M. C. (1990). Learning scientific reasoning skills in microcomputer-based laboratories. *Journal of Research in Science Teaching*, *27*(2), 173–191.

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and Quasi-Experimental Studies of Inquiry-Based Science Teaching: A Meta-Analysis. *Review of Educational Research, 82*(3), 300–329. https://doi.org/10.3102/0034654312457206

Gerard, L. F., & Linn, M. C. (2016a). Using Automated Scores of Student Essays to Support Teacher Guidance in Classroom Inquiry. *Journal of Science Teacher Education*, *27*(1), 111–129. https://doi.org/10.1007/s10972-016-9455-6

Gerard, L. F., & Linn, M. (2016b, April 10). Writing and Revising in Science. Paper presented at the Annual Meeting of the American Educational Research Association.

Gerard, L., & Linn, M. C. (2022). Computer-based guidance to support students' revision of their science explanations. Computers & Education, 176, 104351. https://doi.org/10.1016/j.compedu.2021.104351

Gerard, L., Bichler, S., Bradford, A., Linn, M. C., Steimel, K., & Riordan, B. (2022a). 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* (pp. 1653–1656). International Society of the Learning Sciences.

Gerard, L., Kidron, A., & Linn, M. C. (2019). Guiding collaborative revision of science explanations. *International Journal of Computer-Supported Collaborative Learning*. https://doi.org/10.1007/s11412-019-09298-y

Gerard, L. F., Vitale, J., & Linn, M. C. (2017). *Argument construction to drive inquiry*. Paper presented at the 12th Conference of the European Science Education Research Association (ESERA), Dublin, Ireland.

Gerard, L., Wiley, K., Debarger, A. H., Bichler, S., Bradford, A., & Linn, M. C. (2022b). Self-directed Science Learning During COVID-19 and Beyond. *Journal of Science Education and Technology*, *31*(2), 258–271. https://doi.org/10.1007/s10956-021-09953-w

Gerard, L., Linn, M. C., & Madhok, J. (2016). Examining the Impacts of Annotation and Automated Guidance on Essay Revision and Science Learning. In C. K. Looi, J. L. Polman, U. Cress, & P. Reimann (Eds.), *Transforming Learning, Empowering Learners: The International Conference of the Learning Sciences (ICLS) 2016* (Vol. 1, pp. 394–401). International Society of the Learning Sciences. https://repository.isls.org//handle/1/141

Gerard, L., Matuk, C., McElhaney, K., & Linn, M. C. (2015a). Automated, adaptive guidance for K-12 education. *Educational Research Review*, *15*, 41–58. https://doi.org/10.1016/j.edurev.2015.04.001

Gerard, L. F., Ryoo, K., McElhaney, K., Liu, L., Rafferty, A. N., & Linn, M. C. (2015b). Automated Guidance for Student Inquiry. *Journal of Educational Psychology*, *108*(1), 60–81. https://doi.org/10.1037/edu0000052

Gerard, L., Wiley, K., Bradford, A., King Chen, J., Breitbart, J., & Linn, M. C. (2020). Impact of a Teacher Action Planner Capturing Student Ideas on Customization Decisions. In M. Gresalfi & I. S. Horn (Eds.), The Interdisciplinarity of the Learning Sciences, 14th International Conference of the Learning Sciences (ICLS) 2020 (Vol. 4, pp. 2077-2084). International Society of the Learning Sciences.

Graesser, A. G., McNamara, D. S., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point&Query, AutoTutor, and iSTART. *Educational Psychologist, 40*(4), 225–234.

Hayes, J. R., & Flower, L. S. (1986). Writing Research and the Writer. *American Psychologist*, *41*(10), 1106–1113.

Higgs, J., & Vakhil, S. (2019). It's About Power. *Communications of the ACM*, *62*(3), 31–33.

Höttecke, D., & Allchin, D. (2020). Reconceptualizing nature-of-science education in the age of social media. *Science Education*, *104*(4), 641–666. https://doi.org/10.1002/sce.21575

Ke, F., & Hoadley, C. (2009). Evaluating online learning communities. *Educational Technology Research and Development, 57*(4), 487–510. https://doi.org/10.1007/s11423-009-9120-2

Koedinger, K. R., & Aleven, V. (2016). An Interview Reflection on "Intelligent Tutoring Goes to School in the Big City." *International Journal of Artificial Intelligence in Education, 26*(1), 13–24. https://doi.org/10.1007/s40593-015-0082-8

Koedinger, K. R., Corbett, A. C., & Perfetti, C. (2012). The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science, 36*(5), 757–798.

Könings, K. D., Seidel, T., & van Merriënboer, J. J. G. (2014). Participatory design of learning environments: Integrating perspectives of students, teachers, and designers. *Instructional Science*, *42*(1), 1–9.

Krajcik, J. S., & Mun, K. (2014). Promises and Challenges of Using Learning Technologies to Promote Student Learning of Science. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of Research on Science Education, Volume II* (pp. 337–360). Routledge.

Krist, C. (2020). Examining how classroom communities developed practice-based epistemologies for science through analysis of longitudinal video data. Journal of *Educational Psychology, 112*(3), 420–443. https://doi.org/10.1037/edu0000417

Kubsch, M., Krist, C., & Rosenberg, J. M. (2022). Distributing epistemic functions and tasks—A framework for augmenting human analytic power with machine learning in science education research. *Journal of Research in Science Teaching*, 1-25. https://doi.org/10.1002/tea.21803

Kyza, E. A., & Agesilaou, A. (2022). Investigating the Processes of Teacher and Researcher Empowerment and Learning in Co-design Settings. *Cognition and Instruction, 40*(1), 100–125. https://doi.org/10.1080/07370008.2021.2010213

Kyza, E. A., Michael, G., & Constantinou, C. P. (2007). The rationale, design, and implementation of a web-based inquiry learning environment. *CBLIS Conference Proceedings 2007 Contemporary Perspective on New Technologies in Science and Education*. https://gnosis.library.ucy.ac.cy/handle/7/64727

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society.* Harvard University Press.

Lee, H.-S., Liu, O. L., & Linn, M. C. (2011). Validating Measurement of Knowledge Integration in Science Using Multiple-Choice and Explanation Items. *Applied Measurement in Education, 24*(2), 115–136.

Lieberman, D. A., & Linn, M. C. (1991). Learning to learn revisited: Computers and the development of self-directed learning skills. *Journal of Research on Computing in Education, 23*(3), 373–395.

Linn, M. C. (1995). Designing computer learning environments for engineering and computer science: The Scaffolded Knowledge Integration framework. *Journal of Science Education and Technology, 4*(2), 103–126.

Linn, M. (2003). Technology and science education: Starting points, research programs, and trends. *International Journal of Science Education, 25*(6), 727-758. https://doi.org/10.1080/09500690305017

Linn, M. C., & Clancy, M. J. (1992). The case for case studies of programming problems. *Communications of the ACM, 35*(3), 121–132.

Linn, M. C., & Eylon, B.-S. (2011). *Science Learning and Instruction: Taking Advantage of Technology to Promote Knowledge Integration.* New York: Routledge.

Linn, M. C., & Hsi, S. (2000). *Computers, Teachers, Peers: Science Learning Partners.* Lawrence Erlbaum Associates.

Linn, M. C., Davis, E. A., & Eylon, B.-S. (2004). The Scaffolded Knowledge Integration Framework for Instruction. In M. C. Linn, E. A. Davis, & P. Bell (Eds.), *Internet Environments for Science Education* (pp. 47–72). Routledge.

Linn, M. C., Donnelly-Hermosillo, D., & Gerard, L. F. (in press). Synergies between learning technologies and learning sciences: Promoting equitable secondary science education. In N. Lederman, D. Zeidler, & J. Lederman (Eds.), *Handbook of Research on Science Education: Vol. III*. Routledge Press.

Linn, M. C., Lee, H.-S., Tinker, R., Husic, F., & Chiu, J. L. (2006). Teaching and Assessing Knowledge Integration in Science. *Science, 313*(5790), 1049–1050. https://doi.org/10.1126/science.1131408

Linn, M. C., Gerard, L., Ryoo, K., McElhaney, K., Liu, O. L., & Rafferty, A. N. (2014). Computer-Guided Inquiry to Improve Science Learning. *Science, 344*(6180), 155-156. https://doi.org/10.1126/science.1245980

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated Scoring of Constructed-Response Science Items: Prospects and Obstacles. *Educational Measurement: Issues and Practice*, *33*(2), 19–28. https://doi.org/10.1111/emip.12028

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, *53*(2), 215–233. https://doi.org/10.1002/tea.21299

Lockheed, M. E., & Frakt, S. B. (1984). Sex Equity: Increasing Girls' Use of Computers. *Computing Teacher*, *11*(8), 16–18.

Mandinach, E., Linn, M., Pea, R., & Kurland, M. (1986). The Cognitive Effects of Computer Learning Environments. *Journal of Educational Computing Research, 2*, 409-410. https://doi.org/10.2190/GQ23-EA33-51BM-5HCT\

Matuk, C., & Linn, M. C. (2018). Why and how do middle school students exchange ideas during science inquiry? *International Journal of Computer-Supported Collaborative Learning*, *13*(3), 263–299. https://doi.org/10.1007/s11412-018-9282-1

McBride, E. A., Vitale, J. M., Applebaum, L., & C. Linn, M. (2016). Use of Interactive Computer Models to Promote Integration of Science Concepts Through the Engineering Design Process. In C. K. Looi, J. L. Polman, U. Cress, & P. Reimann (Eds.), *Transforming Learning, Empowering Learners: The International Conference of the Learning Sciences (ICLS) 2016* (Vol. 2, pp. 799–802). International Society of the Learning Sciences. https://repository.isls.org//handle/1/313

McElhaney, K. W., Chang, H.-Y., Chiu, J. L., & Linn, M. C. (2015). Evidence for effective uses of dynamic visualisations in science curriculum materials. *Studies in Science Education, 51*(1), 49–85. https://doi.org/10.1080/03057267.2014.984506

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(2), 57–74. https://doi.org/10.1017/S0140525X10000968

Mokros Ph.D., J. R., & Tinker Ph.D., R. F. (1987). The impact of microcomputer-based labs on children's ability to interpret graphs. *Journal of Research in Science Teaching, 24*(4), 369–383. https://doi.org/10.1002/tea.3660240408

Papert, S. (1980). *Mindstorms: Children, Computers, and Powerful Ideas.* New York: Basic Books.

Puntambekar, S., & Hübscher, R. (2005). Tools for Scaffolding Students in a Complex Learning Environment: What Have We Gained and What Have We Missed? *Educational Psychologist, 40*(1), 1–12.

Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., Kyza, E., Edelson, D., & Soloway, E. (2004). A Scaffolding Design Framework for Software to Support Science Inquiry. *The Journal of the Learning Sciences*, *13*(3), 337–386.

Reiser, B. J., Novak, M., McGill, T. A. W., & Penuel, W. R. (2021). Storyline Units: An Instructional Model to Support Coherence from the Students' Perspective. *Journal of Science Teacher Education, 32*(7), 805–829. https://doi.org/10.1080/1046560X.2021.1884784

Richland, L. E., Linn, M. C., & Bjork, R. A. (2007). Chapter 21: Instruction. In F. T. Durso (Ed.), *Handbook of Applied Cognition* (2nd ed., pp. 555–583). John Wiley & Sons, Ltd.

Riordan, B., Bichler, S., Bradford, A., King Chen, J., Wiley, K., Gerard, L., & C. Linn, M. (2020). An empirical investigation of neural methods for content scoring of science explanations. *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 135–144. https://doi.org/10.18653/v1/2020.bea-1.13

Ryoo, K., & Linn, M. C. (2012). Can dynamic visualizations improve middle school students' understanding of energy in photosynthesis? *Journal of Research in Science Teaching, 49*(2), 218–243. https://doi.org/10.1002/tea.21003

Scardamalia, M., & Bereiter, C. (2006). Knowledge building: Theory, pedagogy, and technology. In K. Sawyer (Ed.), *Cambridge Handbook of the Learning Sciences* (pp. 97–118). Cambridge University Press.

Shear, L., Bell, P., & Linn, M. C. (2004). Partnership Models: The Case of the Deformed Frogs. In M. C. Linn, E. A. Davis, & P. Bell (Eds.), *Internet Environments for Science Education* (pp. 289–314). Mahwah, NJ: Routledge.

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research, 78*(1), 153–189.

Slotta, J. D., & Linn, M. C. (2009). *WISE Science: Web-Based Inquiry in the Classroom.* New York: Teachers College Press.

Smetana, L. K., & Bell, R. L. (2012). Computer Simulations to Support Science Instruction and Learning: A critical review of the literature. *International Journal of Science Education, 34*(9), 1337–1370. https://doi.org/10.1080/09500693.2011.605182

Smith III, J. P., diSessa, A. A., & Roschelle, J. (1994). Misconceptions Reconceived: A Constructivist Analysis of Knowledge in Transition. *Journal of the Learning Sciences, 3*(2), 115–163. https://doi.org/10.1207/s15327809jls0302_1

Sun, D., Looi, C.-K., & Xie, W. (2017). Learning with collaborative inquiry: A science learning environment for secondary students. *Technology, Pedagogy and Education*, *26*(3), 241–263. https://doi.org/10.1080/1475939X.2016.1205509

Tansomboon, C., Gerard, L., & Linn, M. C. (2015). *Impact of knowledge integration and teacher simulated guidance on student learning*. Paper presented at the Annual Meeting of the American Education Research Association, Chicago, IL.

Tansomboon, C., Gerard, L. F., Vitale, J. M., & Linn, M. C. (2017). Designing Automated Guidance to Promote Productive Revision of Science Explanations. *International Journal of Artificial Intelligence in Education, 27*(4), 729–757. https://doi.org/10.1007/s40593-017-0145-0

Toth, E. E., Suthers, D. D., & Lesgold, A. M. (2002). "Mapping to know": The effects of representational guidance and reflective assessment on scientific inquiry. *Science Education, 86*(2), 264–286. https://doi.org/10.1002/sce.10004

VanLehn, K. (2011). The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist, 46*(4), 197–221. http://dx.doi.org/10.1080/00461520.2011.611369

Vitale, J. M., McBride, E., & Linn, M. C. (2016). Distinguishing complex ideas about climate change: Knowledge integration vs. specific guidance. *International Journal of Science Education*, *38*(9), 1548–1569. https://doi.org/10.1080/09500693.2016.1198969

Weber, E. U., & Stern, P. C. (2011). Public understanding of climate change in the United States. *The American Psychologist*, *66*(4), 315–328. https://doi.org/10.1037/a0023253

Wieman, C. E., Adams, W. K., & Perkins, K. K. (2008). PhET: Simulations That Enhance Learning. *Science*, *322*(5902), 682–683.

Wiley, K., Gerard, L., Bradford, A., & Linn, M. C. (in press). Teaching With Technology: Empowering Teachers and Promoting Equity in Science. In A. M. O'Donnell, J. Reeve, & N. Barnes (Eds.), *Oxford Handbook of Educational Psychology*. Oxford University Press.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education*, *56*(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757

Zheng, B., Lawrence, J., Warschauer, M., & Lin, C.-H. (2015). Middle School Students' Writing and Feedback in a Cloud-Based Classroom Environment. *Technology, Knowledge and Learning*, *20*(2), 201–229. https://doi.org/10.1007/s10758-014-9239-z

Zhu, M., Liu, O. L., & Lee, H.-S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, *143*, 103668. https://doi.org/10.1016/j.compedu.2019.103668