Impact of an Adaptive Dialog that Uses Natural Language Processing to Detect Students' Ideas and Guide Knowledge Integration

Abstract

This study leverages Natural Language Processing (NLP) to deepen understanding of how students integrate their ideas about genetic inheritance while engaging in an adaptive dialog. In Study 1, informed by Knowledge Integration (KI) pedagogy, we used responses from 1485 students to test one NLP model to detect the ideas students express when explaining why siblings look similar but not identical and another NLP model to holistically score their response for KI. In Study 2 we used the tested NLP models from Study 1 to design an adaptive dialog that responds to students' detected ideas. We assessed the impact of the dialog on students' level of KI. We embedded the dialog in a web-based unit and implemented it in 5 middle and high schools with 11 teachers and 610 students. Students' KI scores significantly improved across the unit, and from their initial to revised responses in the dialogs. Consistent with KI, students significantly added differing new accurate ideas. They generally linked their vague ideas to new ideas rather than dropping vague ideas. Two patterns emerged: Students who achieve partial KI form links between new accurate and initial vague ideas; Students who progress to integrated KI distinguish between initial vague and accurate ideas plus new accurate ideas to form varied links. These results clarify that students follow multiple paths to combine their ideas and construct coherent responses while studying a unit featuring adaptive dialogs. They point to designs for adaptive guidance to build on students' ideas and promote integrated understanding.

Keywords: Knowledge integration, adaptive guidance, natural language processing, science, dialog, dialogue

Educational Impact And Implications Statement

Contemporary artificial intelligence tools offer promise for guiding diverse students in classrooms to deepen their understanding of science. We exploit Natural Language Processing (NLP) to diagnose each student's science ideas about genetic inheritance. An NLP chatbot dialog was designed in partnership with middle and high school science teachers, embedded in a web-based genetics unit, and implemented in 5 schools.The NLP chatbot uses knowledge integration pedagogy to engage the student in a dialog to integrate their specific ideas about genetic inheritance. The chatbot recognized each student's initial, often vague ideas about genetic inheritance, asked tailored questions, and helped students generate a more coherent and accurate explanation. The findings demonstrate that the chatbot improves students' science understanding. Students liked the chatbot and wanted the conversation about their science ideas to continue.

Introduction

In this research we leverage Natural Language Processing (NLP) to deepen understanding of how students integrate their ideas about genetic inheritance while engaging in an adaptive dialog. Students have many ideas about genetic inheritance often developed from observations of the natural world (siblings look alike), cultural practices (dressing twins in identical outfits), or science classes (children get genetic information from their parents). The Knowledge Integration (KI) pedagogy argues for valuing each student's ideas as efforts to make sense of their experiences. Prior research on KI characterizes the advantage of making science accessible by connecting science topics to students' lived experiences. Prior research also identifies the processes that support students to integrate their views including initially expressing their ideas and then discovering, distinguishing, and making links by sorting out their ideas.

Students can develop integrated knowledge of science by generating and revising explanations. Students generate an explanation using ideas they have developed from their prior experiences, intuitions, and ideas they have gathered during class instruction. Students need guidance to refine their explanations using evidence. Research suggests that effective guidance affirms a student's initial ideas, encourages students to analyze the evidence underlying their ideas to distinguish their accuracy and coherence, and helps students to seek new ideas from instructional resources to clarify their explanations. Effective and responsive guidance is essential to support students to revise and improve their explanations by incorporating new ideas and clarifying the links among their initial and new ideas. Research demonstrates that this process of generating explanations, getting guidance, and revising is challenging for most students and contributes to improved student science understanding (Chi et al., 1994; Gerard et al., 2015a; Gerard & Linn, 2022). Scientists have repeatedly recognized the role of revision as crucial for incorporating both experimental refinements and scientific discoveries. Guiding students to generate and revise explanations is rare however in most secondary science classrooms (Banilower, 2019). Providing guidance that is responsive to students' varied ideas is challenging for teachers who have 31+ students in a class across 5-7 classes. Further, teachers are frequently reassigned to teach new science topics making it difficult for many teachers to anticipate their students' ideas and plan guidance, especially when they are teaching a topic for the first time.

In this research we leverage NLP to engage students in a dialog about their science ideas, prompting students to generate explanations from their intuitions, experiences and instruction and refine their explanations by analyzing the evidence underlying their ideas. In Study 1 we design and test NLP models to detect the distinct ideas within a student response and holistically score the response for KI. In Study 2 we use the NLP models to build an adaptive dialog that responds to the ideas held by the student and we assess gains in their KI level. We embed the dialog in a web-based unit and study its impact on students who are guided by 11 teachers in five middle and high schools. To deepen understanding of KI we analyze how the dialog responds to the varied ideas students express and the patterns students follow as they built on their ideas, discover new ideas, distinguish evidence-based ideas, and make links between ideas.

Knowledge Integration Pedagogy

The instruction, assessment, professional development, and dialog design in this research is informed by constructivist Knowledge Integration (KI) pedagogy (Kali, 2006; Linn, Donnelly, & Gerard, 2023; Linn & Eylon, 2011). The instruction connects to and builds on student ideas to promote integrated understanding of genetic inheritance. To make genetic inheritance accessible (a key tenet of KI), the instruction engages students in investigating their own inherited characteristics and in exploring a case study of a student with cystic fibrosis. The case study also raises social issues by illustrating the trade-off between the right to privacy and the information a student with cystic fibrosis may need to share with others (Morales-Doyle, 2017; Vakil & Higgs, 2019). Making the unit accessible supports students to connect the topic to their own cultural practices and encourages students to express their own ideas, thus reinforcing their identity as a science learner (Bang & Medin, 2010; Rosebery et al., 2010).

Knowledge integration theorizes about the patterns students follow to form coherent scientific explanations, emphasizing that students start a science unit with many vague or incomplete ideas and benefit from well-designed guidance as they sort out their ideas and refine their explanations. Research on instruction that succeeds in guiding students to convert these disparate ideas into coherent arguments reveals that students benefit from considering all of their ideas (diSessa, 1999; Matuk & Linn, 2023), discovering new ideas (Furtak et al., 2012), distinguishing among their ideas (Coll et al., 2005), and reflecting on the relationships among their ideas (Williams et al, 2012). Guiding students to distinguish among their ideas is particularly crucial, has been referred to as accountable talk (Resnick et al., 2018), and is often neglected in science classes (e.g., Smith, diSessa, & Roschelle, 1994). Recent psychology studies show the advantage of interleaving instructed topics, which supports the value of supporting students to distinguish among ideas to improve both immediate and long term retention (Birnbaum et al., 2013; Firth, 2021). Without opportunities to consider all of their ideas and to distinguish among these ideas, students may revert to their initial ideas (e.g., Butler & Roediger, 2008; Clark, 2006).

To deepen understanding of KI, we design NLP models to detect the distinct ideas in students' science explanations, facilitate a responsive dialog, and assess the KI level of their explanations prior to and after they engage in the dialog. The KI rubric assesses student explanations based on the degree to which they link ideas with evidence and rewards students for distinguishing among their initial and new ideas to form coherent, evidence-based arguments (Table 1). We incorporate the NLP models and adaptive dialog into an existing KI Genetic Inheritance unit (Matuk et al., 2019; Obaid et al., 2023). The KI instruction encourages students to make self-directed or autonomous revisions to their response, rather than supplying students with the correct answer (Davis, 2003; Pintrich, & De Groot, 1990). We use the NLP models to analyze how students' idea changes relate to their KI.

By focusing on the fine-grained processes of discovering, distinguishing, and linking ideas that are detected by NLP when students revise their responses, this work extends prior work on KI guidance (e.g. Gerard & Linn, 2022; Gerard et al., 2015b; Williams, 2008). Prior research has demonstrated the value of using the KI framework to design an NLP model that generates a holistic score for the overall coherence and accuracy of a student response and to develop adaptive guidance to help each student move up one level in the rubric (Gerard et al., 2015b; Vitale et al., 2016). In this current research we build on these findings to develop new NLP methods to detect the distinct ideas within a student response, and study how guidance that is responsive to the idea(s) a student expresses impacts the student's generation of new ideas within a dialog and refinement of their initial ideas. We study how students use the ideas they raise in the dialog to revise their explanation and strengthen their knowledge integration.

Generating Responses and Knowledge Integration

Requiring students to generate responses to open-ended problems has proven beneficial. When students generate responses and monitor comprehension during instruction they may reflect on their multiple ideas and discern links among them (Chi et al., 1989; Davis, 2003; Rivard, 1994). In laboratory studies, Chi et al (1989) found that students who spontaneously engaged in what they called self-explanations, outperformed students who did not self-explain by generating significantly more ideas and linking the ideas they generated together. In addition, prompting students to self-explain while reading texts resulted in higher scores due to consideration of more ideas (Chi et al., 1994). This result aligns with the finding that encouraging self-directed learning where students distinguish among their own ideas promotes KI (e.g., Gerard et al., 2022).

Research in both laboratory and classroom studies reinforces the value of generating responses compared to rereading the material (Richland et al., 2007; Richland et al., 2005; Roediger & Karpicke, 2006). Replications and extensions of the impact of generating explanations in classroom studies show the benefit of instructional designs inspired by desirable difficulties (Bjork & Bjork, 2020) and KI (Linn, Donnelly-Hermosillo, & Gerard, 2023). Students who generated self-explanations of photosynthesis rather than reading accurate explanations made greater pretest to posttest learning gains (Ryoo & Linn, 2014). Furthermore, students studying the impact of CO2 pollution on climate demonstrated that KI guidance that elicited explanations had a more sustained impact than guidance providing accurate information, as shown on a delayed posttest (Vitale et al., 2016). These classroom studies promote KI by eliciting student ideas, supporting students to discover more ideas, and then prompting students to generate explanations by recognizing the ideas they hold, distinguishing among their ideas, and combining ideas in a coherent explanation. These studies and multiple replications demonstrate the benefit of engaging students in generating explanations compared to reading, rereading, or reviewing material.

Promoting Knowledge Integration Revisions

How to design guidance to help students strengthen their explanations while studying new topics is a crucial question in science education (Berland et al., 2016; Osborne, 2014). Students make integrated revisions when they respond to science instruction by filling gaps in evidence, resolving inconsistencies among ideas, and strengthening the links between ideas. KI revision involves evaluating the initial explanation in light of new information and reformulating the connections among ideas to increase the accuracy and coherence of a response (Gerard & Linn, 2022). Review studies show that guidance design matters and that adaptive guidance is especially beneficial for guiding revision (Black & Willam, 1998; Hattie & Temperley, 2007; ; Kluger & deNisi, 1996).

Science teachers want to guide students to build on their distinct science ideas (Luna, 2018). In a recent study, three teachers wore a video camera over 13 lessons and recorded each time they noticed a student's science idea while teaching. Researchers categorized the types of ideas teachers noticed. They found that consistently across the three teachers, they most frequently noticed student ideas that were categorized as ‘complex sense-making’. Teachers explained that they captured these moments because they involved students using their observations or experiences to infer why or how something happens. These findings resonate with KI pedagogy and integrated revision. Effective guidance for students involves recognizing their observations and promoting their analysis of the evidence. Prior research converges on the finding that when teachers prompt students to make connections between their ideas and evidence presented within instruction, students demonstrate significantly greater science learning gains compared to when teachers guide students by solely identifying gaps in student knowledge or providing students with correct answers (Ruiz-Primo & Furtak, 2005; Puntembekar, Stylianou & Goldstein, 2007).

Many secondary school science teachers value instructional materials that help them elicit student thinking and interpret the reasoning underlying a student’s idea (Davis & Krajcik, 2005). Due to the frequent reassignment of teachers to new grades or schools and recurring changes in which science topics they are expected to teach, it is often challenging for teachers to anticipate the range of ideas their students may express and provide tailored guidance (Attebury et al., 2017). Furthermore, secondary science teachers in our state appreciate ways to amplify their guidance because they have on average 31 students in each class, and 5-7 class periods, for a total of up to 217 students (NCES, 2018). In this research, we develop NLP-based guidance to assist teachers in promoting KI by guiding students to analyze the evidence underlying their observations, to discover new ideas, to distinguish evidence-supported ideas, and to strengthen the links between ideas.

Using NLP to Design Adaptive Guidance for Integrated Revision

Determining promising designs for adaptive guidance is an active area of research (de Jong et al, 2023; Koedinger & Aleven, 2007; Lee et al., 2021; Vitale, McBride & Linn, 2016). Partnerships of education researchers, computer scientists, and teachers are collaborating to design computer-based adaptive guidance systems using machine learning and NLP specifically (Zhai et al., 2020). In a series of previous classroom studies Gerard, Ryoo et al (2015) demonstrated that adaptive guidance, based on an NLP derived KI score was more effective than generic guidance (e.g. add evidence to your explanation) or specific guidance (e.g. what type of energy enters the plant) and similar to guidance provided by knowledgeable and experienced teachers with limited time. The NLP derived KI score assessed the extent to which the student's science explanation connected accurate ideas and the associated guidance was designed to move the student up one level in the KI rubric. Meta-analyses of studies on computer assigned guidance for student generated explanations and drawings report findings that further corroborate the benefits of guidance aimed at helping students to distinguish among ideas rather than providing the right answer (Gerard, Matuk, et al., 2015; Kluger & DeNisi, 1996; Richland et al., 2005; Vitale et al, 2015).

While studies demonstrated that providing students with NLP-based adaptive guidance is a promising approach to strengthen learning, prior research also highlights the challenges associated with guiding students to use adaptive guidance to make integrated revisions to their explanations. First, research reveals that when students receive adaptive guidance on their written science work, they frequently add disconnected ideas, address grammatical errors, or do not revise (Crawford, Lloyd, & Knoth, 2008; Tansomboon, Gerard, Vitale, & Linn, 2017; Zhu et al., 2020). Many students tack on new ideas to their initial explanation rather than making integrated revisions. Students often then express two disconnected ideas, and in some cases contradictory ideas, in one explanation (Gerard & Linn, 2022). Further, research suggests that when guidance offers students a conflicting view, students often hold onto the two contrasting views simultaneously or defend and strengthen their own view, rather than integrating the views by seeking additional evidence (Kyza, 2009; Mercier & Sperber, 2011; Nickerson, 1998). This may lead students to develop fragmented or superficial knowledge (Campbell, Schwarz, & Windschitl, 2016; Clark, 2006; diSessa & Minstrell, 1998).

In this research, we address these issues by coupling NLP holistic KI scoring with NLP idea detection to identify specific ideas in student responses and engage students in a dialog about their ideas. The idea detection scores are used to design an adaptive dialog that prompts each student to analyze the evidence underlying their idea and to use evidence to distinguish among their ideas. We conjecture that this will support students to strengthen KI as they integrate new ideas with their initial ideas when revising their explanations after the dialog.

Present Study: Overview of Study 1 and Study 2

In Study 1 we describe the development of idea-detection and KI NLP models and evaluate their accuracy and feasibility for classroom use. In Study 2 we embed the NLP models in a web-based inquiry science unit to facilitate an adaptive dialog and assess student KI progress. We use the logged NLP idea-detection scores to explore how students add and drop ideas as they engage in adaptive dialogs, and the NLP KI scores to assess how students' idea changes contribute to their overall understanding.

This work was conducted in partnership with teachers who are a part of a larger research practice partnership in the Western United States created to develop and test NLP-based adaptive dialogs for secondary school science instruction. The partnership includes science teachers, learning scientists, and computer scientists. Teachers were recruited who had interest in using web-based curriculum materials and who worked in public secondary schools serving racially and socio-economically diverse students. Partners participate in periodic workshops, school-based meetings, and university-based meetings where they design the NLP idea-detection rubrics and adaptive dialog prompts. They also customize the web-based curriculum unit to embed the adaptive dialogs in a KI-informed sequence of instruction. The partnership uses the open-source Web-Based Inquiry Science Environment (WISE; Linn et al., 2023). WISE combines technological advances including visualizations, collaborative tools and learning analytics to support teachers in guiding students to investigate contemporary science issues.

Study 1: Evaluation of NLP models for KI and idea detection

We evaluated the accuracy of the NLP KI and idea detection models for a set of training data and for data collected in the classroom when the dialogs were used for instruction. To build automated scoring NLP models, human raters first used a set of scoring rubrics to score students' responses. The training data for the automated scoring system consisted of the text of the responses and the human-assigned scores. We then deployed the models in the classroom. Human raters scored a subset of the data to establish the accuracy of the NLP models for the new classrooms. We addressed two questions:

1. How accurate are the KI and Idea Detection NLP models from a machine learning perspective when evaluated with training data?
2. How accurate are the KI and Idea Detection NLP models when evaluated in the context of a new classroom setting?

Background on NLP Methods

We used a holistic KI rubric and an analytic idea rubric to build the NLP models. The holistic KI rubric required the human rater to assign one score, simultaneously considering all criteria from the scoring rubric (Table 1). The analytic idea rubric involved scoring each idea in a student response (Table 2). Most research studies employ holistic rubrics (Shermis, 2015; Institute of Education Sciences, 2022, 2023) and specifically in science assessment, holistic scoring has been widely used (Kaldaras et al., 2021; Shermis, 2015; Zesch et al., 2023; Zhai et al., 2020). Because student responses in science assessments are often relatively short, holistic rubrics are feasible.

However, holistic rubrics have limitations. Because holistic scoring systems are trained on a single score per response, they can erroneously leverage spurious correlations between response characteristics and scores (e.g., Ding et al., 2020; Filighera et al, 2020). This can raise validity concerns (e.g., Myers & Wilson, 2022; Kaldaras & Haudek, 2022). Although both holistic rubrics and analytic rubrics are designed to score student responses on an ordinal scale, when automatically holistically scored assessments have been used for formative assessment in the classroom, researchers report the limitations of the summary scores they generate (Lee et al, 2021; Puntembekar et al., 2023; Rafferty et al., 2013). To address these limitations of holistic scores and to align with the KI pedagogy that focuses on student ideas, we and others began to explore directly *detecting ideas* in responses (Riordan et al., 2020; Schulz et al., 2019). Idea detection in educational assessment attempts to find and label “spans” of words in responses that correspond to ideas in the scoring rubric.

Methods

“Siblings” Constructed Response Prompt

We developed NLP KI and idea detection models for a constructed response item referred to as *Siblings*. The Siblings item, *Why do siblings look similar but not exactly the same?* (Figure 1) *,* was developed in prior research to elicit links among student ideas about the inheritance of genes from each biological parent and genetic variation (Table 1: KI Rubric) (Obaid et al., 2023). It encouraged students to draw on their personal experience or observations of siblings and parents’ appearances and ideas from the WISE genetics unit to form a response.

NLP Idea Detection and KI Modeling

Training data. A corpus of 1485 student responses for the Siblings item was collected from 2016-2018 in prior research with schools whose school demographics[1] are similar to the schools that would use the NLP models in Study 2, to build the NLP models. Schools in the training data are in the Western United States.

Idea Detection rubric. Two teachers, one a former high school science teacher who is now a Learning Sciences PhD student, and one current high school biology teacher who is a partner in the research project, analyzed 10% of the student responses in the training data to create an idea rubric (Table 2). An idea was typically made up of a single sentence, or a phrase within a sentence within a response (e.g. underlined segments in Table 2). The teachers set the boundaries for an idea by identifying expressions within an explanation that they would typically respond to in the classroom. This meant teachers identified distinct ideas in each student's explanation that they would want to build on or probe to help the student deepen their understanding of genetic inheritance. One student response could include zero, one, or more ideas[2].

The initial idea rubric enumerated the distinct ideas emergent from the students' answers in the training data grouped in terms of accurate ideas; vague ideas; and off-task ideas. This rubric was reviewed by research partners, resulting in changes such as merging idea categories with substantial overlap, and elaborating idea category criteria. The two teacher raters reapplied the updated rubric to the 10% data sample to ensure the refined version captured all expressed student ideas. The final version of the Siblings idea rubric identified 25 distinct ideas that were found in students' responses.

---

[1] Student demographics for training data: Schools range from 94%-37% of students who identify as non-White; 93%-12% students eligible for a free or reduced price lunch; 32%-2% of students designated as English Language Learners. Compare to student demographics for schools in Study 2 who used the models (Table 4).

[2] Relationships between ideas were not annotated.

KI rubric. Researchers also used a KI scoring rubric (Table 1) to assess the overall quality of each student's response to Siblings, capturing the degree to which students connected accurate ideas. The range of possible scores in the KI rubric are from 1 (off task) to 5 (linking two or more accurate ideas). The Siblings rubric was developed in prior research (Harrison et al., 2018; Obaid et al, 2023).

Annotation. Human raters assigned a KI score to each response, and annotated the ideas in responses, using the INCEpTION platform (Klie et al., 2018). This was a two-level annotation process. We computed a kappa for the rater agreement on the KI score for the student's explanation, and a second kappa for the rater agreement on the ideas present/absent in the student's explanation. Inter-rater reliability for idea-detection was calculated based on the agreement for the presence/absence of an idea and not on the length span of the annotated idea segment. The raters trained on the data until they achieved inter-rater reliability for scoring of the KI scores above a Cohen's Kappa of .85, and scoring of the idea-detection above a Cohen's Kappa of .85. Once reaching inter-rater reliability for KI and idea detection each researcher annotated 50% of the student responses. They assigned a KI score for each explanation, and a tag for each distinct idea within each response.

NLP Modeling for idea detection. The computer scientist used the teacher-annotated training data to build the NLP model for idea detection. This required tackling two interrelated NLP problems: (1) classifying a span of text into a set of known classes (e.g., a set of pre-identified categories of student ideas); and (2) detecting the boundaries of the idea within the span of text (segmentation). To address these challenges, the NLP model used a sequence labeling or token classification approach for idea detection (Riordan et al., 2020; Schulz et al., 2019). That is, for each word in a student response, a model assigned one or more idea categories. A contiguous span of these predictions formed a predicted idea's *span*. Since individual words could belong to more than one idea category, idea spans from different categories could overlap. Assigning idea categories to words is a multi-label classification task: for each word, the model makes multiple binary classification predictions about whether the word belongs to an idea category's span.

Pre-trained transformer models (Devlin, Chang, Lee, & Toutanova, 2018) were used for both holistic scoring and idea detection. Transformer model training involved starting from a large pre-trained model and fine-tuning the model weights on training examples specific to holistic scoring or idea detection (Alhindi & Ghosh, 2021). See Appendix NLP Modeling for details about the model architectures and training regimes.

The multi-label idea detection model was trained and validated with 10-fold cross validation for hyperparameter tuning and model selection. When the NLP model achieved sufficient accuracy for the designated idea categories [see Results] (Schulz et al, 2019), the model and adaptive dialog were embedded into the WISE Genetics unit.

NLP Modeling for KI. In this work, we modeled the holistic scoring task as a text regression problem: for each student response, we assigned an ordinal score. Leveraging ordinal scores to train the models can improve human-machine agreement compared with using unordered class labels (e.g., Jescovitch et al., 2020). Moreover, we employed instance-based machine learning models, which learn from examples in the form {text, score} (Horbach & Zesch, 2019). While similarity-based methods where individual responses are compared with a small set of reference responses can be competitive for holistic scoring tasks, instance-based approaches typically yield higher accuracy (Bexte, Horbach, & Zesch, 2022).

Classroom data. We embedded the idea detection and KI NLP models in the WISE Genetics unit to facilitate and assess an adaptive dialog. Teachers in five schools with demographics similar to those in the training data, and located in the same geographic region as schools in the training data, implemented the unit with their 610 students (See Table 4 for participant details). To evaluate the accuracy of the NLP models with the new classroom data, we used *R* to create 5 random samples of the classroom data, each with 100 student responses to Siblings (v4.2.1, R Core Team, 2022). We selected the sample with the greatest distribution across the 5 schools, KI scores, and the least missing data to evaluate.

Data Analysis

RQ1: Evaluation of Training Data

We used 10-fold cross validation to evaluate the human-machine agreement of the KI and idea detection NLP models. The KI scoring models were evaluated on quadratic weighted kappa (QWK). QWK is a measure of agreement that ranges between 0 and 1. QWK is the industry standard for evaluation of machine and human scoring in education research (Nehm et al., 2012; Zhai et al., 2022) because it compensates for randomness or chance in inter-rater agreements (Fleiss & Cohen, 1973; Haberman, 2019). We used Nehm et al (2012) criteria to evaluate the accuracy of the QWK for classroom use, an accepted criteria in science education research on machine learning (e.g. Zhai et al., 2022).

For the idea detection model we evaluated human-machine idea agreement using a response-level micro-averaged F1 score. The F1 score is a combined measure of precision and recall across each response in the data. The macro-average weights all idea categories equally while the micro-average is affected by the frequency with which an idea occurs. We evaluated at the response-level because the NLP models use the presence/absence of an idea in a response to assign guidance in a dialog. There is currently no agreed-upon threshold for idea detection accuracy for model deployment in classrooms (Schulz et al., 2019).

RQ2: Evaluation of Classroom test data

The classroom data was loaded into INCEpTION and one of the teachers who annotated the training data annotated the classroom data using the same idea rubric. The same metrics were used to evaluate the classroom data as were used to evaluate the training data.

Transparency and Openness

For both Study 1 and Study 2, all curriculum materials including the embedded NLP models and the coding rubrics are available at https://wise.berkeley.edu under a Creative Commons license. Due to our approved IRB's Human Subjects protocol for conducting research with a vulnerable population (children under 18) the study data are not made publicly available. Authors will provide an explanation of the R syntax and how to adapt it for others' learning analytics data upon request. This study was not pre registered.

Results

RQ1: Training Data

The evaluation of the training data demonstrated a high level of agreement between the NLP model and human scoring for KI with a QWK of .85. Values between .81-1.00 are considered to be "almost perfect" in evaluation of machine learning in education (Nehm et al., 2012; Zhai et al., 2022).

Similarly, the evaluation demonstrated a reasonable level of agreement between the NLP idea detection and teacher detection of student ideas within students' responses. The overall F1 score achieved by the idea-detection model was 0.78. The model demonstrated reasonable precision (0.77) and recall (0.78) in predicting idea spans.

We also computed the F1 score for each of the idea categories (Table 3). Four student idea categories had an F1 score of 0 due to low frequency in the training data (below 35 annotations). Ideas 11 (same gender children share more traits), 12 (siblings have different chromosomes; environmental factors), 13 (meiosis), and 22 (cell division). As a partial mitigation for algorithmic bias, these idea categories were removed from the final NLP model prior to embedding it in instruction. This resulted in 17 distinct student idea categories that could be detected reliably by the model, including 9 vague ideas and 8 accurate ideas. Fifteen of these 17 distinct ideas had F1 scores above .5, meaning these idea categories were detected reasonably well. Two of the ideas (2 & 19) had F1 scores in the .3-.4 range, meaning it was challenging to detect these ideas, largely due to the low frequency of annotations in the training data.

RQ2: Classroom Data

The evaluation of the human-machine agreement for KI scoring of the classroom data resulted in a QWK of .65, a substantial level of agreement (Nehm et al., 2012; Zhai et al., 2022).

Agreement between the NLP idea detection and teacher detection of student ideas within students' responses in the new classroom data (Appendix: Table 1) was reasonable. The overall F1 score achieved by the idea-detection model was .67; the model demonstrated reasonable precision (0.63) and recall (0.73) in predicting idea spans. The F1 score for each of the idea categories was similar to the training data with three of the ideas (4, 19, 23) having F1 scores in the .3-.4 range, meaning it was challenging to detect these ideas. Two ideas were not present in the data sample (2, 25) meaning they lack an F1 score.

Study 1 Discussion

The idea-detection and KI NLP models demonstrated acceptable results in detecting students' accurate and vague ideas about genetic inheritance and the integration of these ideas, based on the F-score and QWK evaluation metrics. Consistent with prior research on NLP modeling, the ideas with a higher frequency of annotation in the training data were more accurately detected (Schulz et al., 2019). While there is no agreed-upon threshold for sufficient idea detection accuracy for model deployment, the F1 scores achieved by our trained model are in line with the accuracy attained by similar models for challenging idea and reasoning detection tasks in the education domain (Schulz et al., 2019).

Developing machine learning models from training data with uneven instances of responses is an active area of research (Johnson & Khoshgoftaar, 2019; *inter alia)*. Various data augmentation and rebalancing strategies have been tested (Lee et al., 2023). At the same time, care is required to ensure that augmented datasets do not reinforce or introduce bias (Gupta et al., 2023).

The evaluation of the models with the classroom data from five secondary schools demonstrates the ecological validity of the KI and idea detection NLP models for science classrooms. The NLP models built using training data from one set of schools were able to detect ideas expressed by new students during regular classroom instruction with substantial accuracy. This suggests that the process for collecting training data from schools in the same geographic region in different years and with different students is sufficient for providing adaptive guidance for new students.

Nevertheless, caution is advised in scaling the NLP idea detection models to new school contexts. There is variation that arises between training data and classroom contexts particularly in terms of the students' background knowledge and their curricular experiences, both of which students integrate to explain science topics. Because of this variation there will always be student ideas that the model is not trained to detect. To address this variation, as well as the four ideas that were dropped from the model due to sparsity in the training data, we adapted a generic dialog prompt for use in Study 2 which is assigned if the NLP model cannot detect a student's idea. The generic prompt is adapted from previously researched KI guidance and aligned with the general character of the student's answer, not to the specific idea they expressed (Gerard et al, 2015). It encourages reflection on one's response, and directs the students to examine evidence related to the question within the curriculum unit.

This NLP method presents novel contributions to AI based assessment of student learning by demonstrating how idea detection models can provide a complementary view of student learning to holistic scoring (Zhai et al., 2020). We used a partnership model of designing an idea rubric with classroom teachers using student responses from schools with similar demographics to those students who would use the models. This partnership process resulted in a rubric with a set of ideas that were emergent from students' experiences and observations in genetic inheritance. Prior research demonstrates that science teachers noticed student ideas that came from their students' own observations and experiences to explain phenomena, rather than noticing normative student ideas alone (Luna, 2018). This idea detection approach aligns with the KI pedagogy of affirming student ideas and supports teachers by recognizing and responding to the student ideas teachers value.

Study 2: How does an NLP idea-detection dialog in a web-based inquiry unit facilitate student knowledge integration?

The results from Study 1 enable the investigation of an adaptive dialog that uses the NLP generated idea scores to help students build on, elaborate and refine their ideas to progress in KI. In Study 2, we used the logged NLP idea detection data to understand the changes students made to their ideas as they engaged in adaptive dialogs. We used the logged KI scores to examine the impact of students' idea changes on their development of coherent understanding.

Research Questions

1. How do students progress in KI as they study a web-based unit and engage in adaptive dialogs in genetic inheritance? What off-task, vague, and accurate ideas do students discover, distinguish among, and integrate that contribute to KI progress?
2. What is the relationship between students' idea changes about genetic inheritance and their progress in KI?

Adaptive Dialog and Genetic Inheritance Curriculum

In this research we developed an adaptive dialog for *Siblings,* which is embedded in a 7-10 day web-based inquiry science unit on genetic inheritance. Students encountered an image of siblings and the Siblings question in the unit (why do siblings look similar but not identical). They chose a “thought buddy” from among a group of images that look like students or robots. The student-selected thought buddy then greeted the student by name and asked them to share their ideas about the Siblings prompt. The student wrote their response. NLP was used to detect the specific ideas in the student’s explanation, and the overall KI quality of their explanation (See Study 1). Based on the idea detected, adaptive guidance was assigned. The student wrote a response and the NLP idea-detection analyzed the student response. Students received a second adaptive prompt to refine the idea detected in their response. After the student responded to the adaptive prompt, the thought buddy prompted the student to use what they learned to revise their initial response. The revised response was scored by the NLP models for idea-detection and KI.

For example, in the dialog shown in Figure 1, the student gave an initial response. The NLP models detected one vague idea (siblings have different genes) and assigned a KI score of 2. The buddy delivered the first adaptive prompt. The prompt elicited more of the student's reasoning to help them refine their detected idea. The student expressed a new accurate idea in response. The NLP idea-detection identified the idea (different combinations of alleles) and assigned the adaptive prompt for that idea to further probe for refinement. The student refined their initial idea in response and added another accurate idea (genotype affects phenotype). The dialog prompted the student to revise their initial response. When revising, the student linked an accurate idea they raised in the dialog to elaborate their initial vague idea: *“Siblings look similar to each other in some ways but not exactly the same because the <u>alleles can pair up differently</u>*

which makes *<u>different DNA sequence</u>.”* The NLP assigned the revised response a KI 3 for linking an accurate idea (different combinations of alleles) to a vague idea (different DNA).

Adaptive Dialog Design

In this investigation the partnership designed and tested an adaptive dialog to improve students’ ability to evaluate their own explanation and make integrated revisions. We built on prior research to design the dialog and the guidance in the dialog, and testing of the dialog guidance for bias among monolingual and multilingual students.

Developing the Dialog Design

We started by extending research on an Annotator showing that annotating a fictional peer’s explanation can help students to improve their explanations. In prior research Gerard & Linn (2022) developed the Annotator to model for students the process of using adaptive KI guidance to make integrated revisions to their science explanations. The Annotator prompted students to help a fictional peer, Mary, use personalized guidance to revise her science explanation. The student was given Mary’s explanation (which has gaps and inaccurate ideas) and three guidance prompts Mary received from the computer. The student selected a guidance prompt and dragged/dropped the guidance onto the part of Mary’s explanation, to indicate where Mary should make the specific revision associated with the prompt (either to add a new idea that was missing or modify an inaccurate idea). The student repeated this process for each of the three guidance prompts. Then, students received one round of personalized KI guidance for their own explanation and were prompted to use the guidance to revise. Students in the control condition received two rounds of personalized KI guidance for their explanation and no Annotator. In both conditions the personalized KI guidance included a hint and a direction to relevant evidence in the unit.

The Annotator condition supported students who initially demonstrated vague ideas in their explanations to make more integrated revisions on an embedded assessment, compared to students in the control condition. This was surprising in that students in the control condition, particularly students who initially expressed vague ideas, had more opportunities to gather new information from the two personalized KI hints they received, compared to students in the Annotator condition who received one personalized KI hint. On the embedded assessment and a posttest explanation revision activity the students who had used the Annotator during instruction made greater improvements to their explanations than the control students. From a KI perspective, these findings suggest that the students who initially expressed a vague idea had many ideas in their existing repertoire. They needed support to distinguish which of these ideas were relevant and how to integrate them when revising. Analyzing the distinct ideas within an explanation supported the students to distinguish which idea from their repertoire was missing in the fictional peer’s explanation and in their own explanation.

The Annotator study demonstrated that analyzing distinct ideas in an explanation with guiding prompts, helped students to identify a gap in an explanation and to select an idea from their repertoire to fill the gap. Annotating was more effective in promoting integrated revision than providing students with multiple, additional new relevant ideas to consider. Students however, selected both inaccurate and accurate ideas, rather than distinguishing accurate from inaccurate ideas when filling a gap in their explanation. This aligns with research suggesting that a dialogue can model the process of identifying gaps in reasoning, compared to a monologue or static instructional materials (Chi et al., 2009; Chi et al., 2017; Van Lehn, 2011).

The adaptive dialog structure used in this study emerged at a workshop with our teacher partners where we first reviewed the Annotator findings. We built on the idea of assessment conversations, used by teachers to help students reflect on their reasoning and connect to a new idea (Ruiz-Primo & Furtak, 2007). The teachers thought that a dialogue with a thought buddy would engage students and that students would be likely to tell the thought buddy ideas that they might hesitate to share with a teacher or peer for fear of being wrong.

We also built on research in NLP and computer-based adaptive dialogs. A review of Intelligent Tutoring Systems used in college laboratory investigations found that student dialogue with a computer tutor was nearly as effective as a dialogue with a human tutor in helping students solve problems in STEM, with an average effect size of .7 for both the human and automated tutor (VanLehn, 2011). Authors conjectured that step-based tutoring from a human or a computer supported students to generate reasoning for each idea spawning generation of new ideas, and encouraged reflecting and refining reasoning at the idea level. By guiding each idea, rather than giving guidance on a student's final answer, the tutor supported the learner to iteratively refine each idea as they constructed an explanation, whereas guiding a student's final answer can present too big of a leap resulting in limited or no refinements (Chi & Wylie, 2014). Similarly, the dialog in this study detected and responded to a distinct idea within each student response. To strengthen KI, the prompts in this study aimed to help students refine their responses by finding gaps in their own reasoning.

Testing Dialog Generalizability

To ensure the generalizability of the adaptive guidance, in prior classroom research, we tested the format of the dialog for bias towards mono- or multi- lingual students (Holtman et al., 2023). We divided the dialog data from 1036 6th-9th grade students into two groups: data from those who reported speaking only English at home (monolingual), and data from those who reported speaking a language other than English at home (multilingual). The mono- and multi-lingual groups expressed distinct ideas across four adaptive dialogs on different secondary school aligned science topics. In the adaptive dialogs, therefore, students received prompts aligned with their distinct ideas. The results indicated that both groups significantly improved their KI score from their initial response to their revised response after the dialog. There were no significant differences in gains between the two language groups. Thus, the NLP idea-detection model development and guidance design displayed no bias in favor of monolingual or multilingual students.

Methods

Participants

Nine teachers with a total of 610 middle and high school students [21% of 9th grade students; 35% 8th; 44% 6th] taught the WISE Genetics unit with the Siblings dialog. As shown in Table 4, schools reflected a range of student demographics ensuring the dialog was designed to benefit diverse populations. The study was approved by University[X]'s Institutional Review Board for research with human subjects.

Developing Adaptive Guidance in the Dialog

We built on the findings from the Annotator (Gerard & Linn, 2022) and Chi et al. 2017 that students often select new ideas during dialogic interactions yet do not necessarily select better ideas. The teachers who developed the idea rubric for the Siblings item and annotated the training data (see Study 1) designed the dialog guidance for each idea. They built on guidance practices they honed while circulating around the classroom and prompting students to refine their explanations. The prompts intended to focus students on using evidence to distinguish among their ideas and to select an evidence-based idea(s). Each prompt targeted a distinct idea in the students' response (See Table 2).

The guidance prompts were reviewed by the members of the research practice partnership and refined (see Sample Guidance in Table 2). This ensured that the guidance connected to the language teachers have heard students use in their classrooms when learning about genetic inheritance and that the questions were clear to secondary school students. For example, the prompt 'Why do siblings look more alike than people who are not related?' was refined to ‘ How does genetic material make siblings look more alike than people who are not related?’ recognizing that students may respond to the initial version by expressing environmental or non-genetic reasons.

When multiple ideas were detected in a student response, the rule-based adaptive guidance was tailored to respond to one of the ideas; we prioritized the accurate idea that the student could best build on to generate a coherent response.

Data

The Siblings item was embedded at three points: Pretest, Within the Unit, and Posttest (Fig. 2).

Pretest

Students completed the Siblings item on the unit pretest in an open-response format without a dialog one day prior to starting the Genetics unit.

Within-unit Adaptive Dialog

Students completed the Siblings item with adaptive dialog within the unit two days after completing the pretest. The adaptive dialog was embedded within the unit after students completed a lesson about genetic inheritance and punnett squares in which they explore dynamic models. The dialog was placed here to engage students in KI using the new ideas they learned about genetic inheritance and their initial ideas.

Posttest Adaptive Dialog

Students completed the Siblings item with adaptive dialog again on the unit posttest after they completed the unit. This was approximately twelve days after they completed the within-unit dialog. Between the within-unit dialog and the posttest dialog students completed personally relevant activities about genetics.

In the pretest, within-unit and posttest adaptive dialogs, the initial prompt asking the Siblings item was the same. Students responded to different adaptive prompts within each dialog based on the ideas detected in their answers.[3] The last prompt in the within unit and posttest dialog was the same, asking students to revise their initial response.

Logged Analytics

WISE logged each student response to the pretest, and each student's interactions in the within-unit and post test dialogs including their initial response, two responses to the prompts in the dialog, and their revised response after the dialog. For every student response, WISE recorded the NLP detected ideas and the NLP generated KI score. The scores for each response were exported into anonymized csv data files for analysis.

Student Interviews

We interviewed a convenience sample of 27 students across four of the schools. The sample included students who completed the within-unit dialog with at least 5-minutes remaining in their class time during a classroom observation. When a student finished the dialog, a researcher asked the student if they would participate in a 5-10 minute interview during the remainder of the class period. The goal of these brief interviews was to collect students' reactions to the dialog design to inform refinements. Questions included: What was it like to interact with your thought buddy? Did you feel that the thought buddy responded to your ideas or no, not really? How did talking to your thought buddy compare to other feedback you sometimes get in science class such as in talking to a peer or teacher?

Data Analysis

For all analyses of student learning, we used the software *R* (v4.2.1, R Core Team, 2022).

[3] We identified 64 student responses that did not receive an idea score in the dialog (labeled "NA" for idea-detection). This was likely due to a technical issue in that the WISE server did not connect to the automated scoring server due to a brief outage in WIFI. Although they did not receive an idea score, they did receive guidance that was pre-authored for KI level 1 responses. In order to determine the actual KI and idea scores for the analysis we ran these student responses through the idea detection model Batch Scorer (https://wise-research.berkeley.edu/class/batchScoreCRater.php) and added the NLP generated idea detection and KI scores. Because these students generated a response and received dialog guidance, we re-incorporated this data into the analyses.

RQ1: Student KI Gains and Idea Changes in Dialog and Across Unit

KI gains. To analyze differences in KI scores between students' Siblings (a) pretest and revised posttest response; (b) initial and revised response in the within-unit dialog; and (c) initial to revised response in the posttest dialog, we conducted paired non-parametric Wilcoxon signed rank tests because the KI score is an ordinal, five-level scale.

Idea changes. We tested if the ideas were mentioned significantly more, or less, frequently between the different timepoints using paired Wilcoxon signed-rank tests for idea groups and Cochran's Q post-hoc McNemar test for single ideas. This Cochran's Q post-hoc McNemar test is used to determine whether there is a statistically significant difference in the distribution of a dichotomous (binary) outcome (in our case: idea mentioned or not) across three or more related groups (in our case: time points).

RQ2: Relationship Between Student Idea Changes and KI

Regression analysis. To examine the relationship between students' idea changes about genetic inheritance and their progress in KI, we regressed students' KI change from one time point to the next time point across the revision activities, on changes in students' ideas between these time points (building on the core idea of the change-score approach). Changes in students' ideas indicated how many ideas, more or less, a student expressed in the subsequent time point compared to the previous time point. We examined this in total and separately for idea groups [off-task, vague, and accurate ideas]. We checked the results of these analyses by regressing the KI Score on the ideas from the previous time point while controlling for the respective KI Score, which yielded similar patterns of results. For clarity, we present the results of the first approach.

Idea changes for partial versus integrated KI. We analyzed the idea changes for the subgroups of students who progressed from KI 1 or KI 2 to a KI 3 in their revised responses, versus those who progressed to displaying links among two or three accurate ideas, KI 4 or KI

5, in their revised responses. We examined each subgroups' ideas in the within-unit and posttest dialogs and across the unit from pretest to the posttest, and identified the ideas that were significantly added or dropped across the revision activities for the subgroup.

Student interviews. We transcribed each interview and identified emergent themes characterizing students' reactions to the dialog design. Themes included: Responds to *my* ideas; Like a more knowledgeable other; Scripted/Redundant; Confusing; Ended too soon. We coded each interview for the presence/absence of each theme.

Results

RQ1: Student KI Progress and Idea Changes in the Dialog and Across the Unit

We analyzed the KI score changes and the accompanying idea changes: (a) across the genetics unit, examining the student's Siblings pretest response to their revised response after the posttest dialog; and (b) across the within-unit and the posttest adaptive dialogs, comparing the student's initial response at the start of the dialog to their revised response immediately after the dialog. We report the overall change in the KI level of the student's response (Table 5). Following KI pedagogy, we then report the changes in the number of ideas and type of ideas students included in their responses, evidencing students' integration of new, accurate ideas (Figure 3; Appendix Tables 2, 3).

Impact of Genetics Unit Across Pretest to Revised Posttest

KI Gain. Students made significant improvement in their KI scores, indicating that they integrated their understanding of genetic inheritance. Specifically, the students' KI scores improved significantly from their pretest response to their revised posttest responses, tested with a paired Wilcoxon signed rank test (W = 6,314, p < .01).

Idea Changes. Taking advantage of the NLP idea detection provided detailed insights into the specific ideas students discovered and integrated across the unit to make these significant improvements. From the pretest response to the revised posttest response, the percent of students who expressed only vague or off-task ideas (KI 1, 2) decreased from 33% to 19%, whereas the percent of students who linked a vague and accurate idea (KI 3) increased from 54% to 63%, and the percentage of students who linked two or more accurate ideas (KI 4, 5) increased from 13% to 18%.

Students significantly increased the total number of ideas they expressed from 2.16 to 2.67 from pretest to posttest demonstrating the integration of new ideas, particularly accurate ideas. Students expressed six of the eight accurate ideas significantly more often in the posttest than in the pretest (Appendix Table 3). They significantly added three accurate ideas that provided a mechanism for genetic variation [*alleles combinations; chance in the genes inherited;* and *Punnett Square].* In addition, they added two ideas about how genes dictate traits [*genotype affects phenotype*; *dominant/recessive genotypes*]. The accurate idea that '*genes are inherited from parents*' was the most common idea mentioned by about 44% of the students in the pretest and 51% of the students in the revised answer of the posttest, suggesting that this was a key building block for understanding across the unit.

Besides significantly adding accurate ideas that provide a mechanism for genetic variation or depict how genes dictate traits, students distinguished vague ideas. Most students clarified their vague ideas including '*siblings have different DNA*' (28% at pretest, 31% at revised posttest) and '*siblings have the same DNA*' (26% at pretest and 27% at posttest) by connecting them to accurate ideas they added during the dialog, rather than dropping these ideas.

Impact of the Within-Unit Dialog

KI Gain. Overall, the success of the dialog was shown in students' constructing significantly more integrated responses about genetics from their initial explanations to their revised explanations after they completed the within-unit dialog. Students significantly improved the KI level of their Siblings responses reflected in the non-parametric Wilcoxon signed-rank test (W = 5,090, p < .01). The mean KI score of students' initial responses was M = 2.66 (SD = .62), slightly lower than their pretest responses possibly due to the more informal format of the Siblings item when embedded within a dialog compared to when in an open-response format on a pretest. The mean KI score of their revised responses at the end of the dialog was M = 3.04 (SD = .78).

Idea Changes. The NLP idea detection shows the specific ideas students discovered and integrated to make this KI progress across the within-unit dialog. From their initial responses at the start of the dialog to their revised responses at the end of the dialog, the percent of students who expressed only vague or off-task ideas (KI 1,2) decreased from 33% to 20% whereas the percent of students who partially linked vague and accurate ideas (KI 3) remained constant (64 % to 61%). The percentage of students who fully linked two or more accurate ideas (KI 4,5) increased from 4% to 18%.

Students significantly increased the number of scientific ideas they expressed from the start to the end of the dialog demonstrating the integration of new ideas, particularly accurate ideas. Students expressed 1.76 ideas in their initial dialog responses and 2.43 ideas in their revised responses. They raised 2.5 ideas, on average, *during* the dialog, consistent with the notion that the dialog supported students to consider new ideas and distinguish among their initial and new ideas to refine their revised response. More specifically, students expressed seven of the eight scientifically accurate ideas significantly more often in their revised responses compared to in their initial responses. The two ideas that increased the most from initial to revised response, focused on how genes dictate traits [*genotype affects phenotype*; *dominant/recessive genotypes*]. In addition, significantly fewer students '*repeated the question as their answer',* suggesting that the within-unit dialog helped students recognize their scientific ideas about genetic inheritance. Specifically, students elaborated rather than repeating the question by significantly adding the vague yet more helpful idea that '*siblings share the same* *genes',* between the initial and revised responses. Although not changing significantly, the constancy in student expression of the idea '*siblings have different genes*' in the initial and revised responses, suggests that this was an alternative path for students who refined their initial vague idea into a more elaborated view of genetics.

Impact of the Posttest Dialog

The KI level and idea changes across the posttest dialog followed the pattern observed for the within-unit dialog.

KI Gain. On the posttest dialog, students made significant improvements in KI level from their initial to the revised Siblings responses as indicated by the significant non-parametric Wilcoxon signed-rank test (W = 4,296, p < .01). Comparing the within-unit dialog responses and the posttest dialog responses, students demonstrated a similar level of KI on their initial and revised posttest responses. The mean KI score of students' initial posttest response was M = 2.62 (SD = .68) and for the revised response was M = 3.06 (SD = .79).

Idea Changes. As in the within-unit dialog, the NLP idea detection showed the ideas students integrated to strengthen their KI across the posttest dialog. Similar to the within-unit dialog, from their initial to their revised posttest response, the percent of students expressing only vague or off-task ideas (KI 1-2) decreased from 37% to 20%; partially linking one accurate idea and a vague idea (KI 3) remained constant from 60% to 63%, and the percentage of students fully linking two or more accurate ideas (KI 4-5) increased from 4% to 18%.

Relatedly, in both the within-unit and posttest dialogs, the adaptive prompting supported students to raise new accurate ideas. Students expressed 1.83 ideas in their initial posttest dialog responses and 2.67 ideas in their revised responses after the posttest dialog. Further, students expressed 2.6 ideas, on average, *during* the dialog, consistent with the notion that the dialog supported students to consider new ideas and distinguish which new idea to incorporate into their revised response. The parallels between the idea changes detected in the within-unit dialog and the post-unit dialog, reinforced the value of the dialog for supporting students in the process of discovering, distinguishing and integrating ideas.

The students' paths from initial to revised responses in the posttest dialog varied, as was reported for the within-unit dialog. Students significantly increased their expression of all but one of the detectable accurate ideas from their initial responses to their revised responses in the posttest dialog. They integrated new ideas related to how genes dictate traits [*genotype affects phenotype*; *dominant/recessive genotypes*] and how variation in inheritance impacts outcomes [*chance in which genes are inherited*; *different combinations of alleles*].

Students also increased their use of several vague ideas. They significantly added the same vague idea as students added in the within-unit dialog,‘*siblings share the same genes*’, again suggesting that this intuitive, vague idea may have been a building block to linking to accurate ideas to elaborate this view. Students also significantly increased expression of the vague idea, ‘*uneven inheritance from each parent*’ from their initial posttest to their revised posttest response; this non-normative idea is common among secondary school aged students (Williams et al, 2011). The idea that ‘*genes are inherited from parents*’ (51% of revised responses) was the most frequently expressed idea in students’ initial and revised posttest dialog responses suggesting that this may have served as a starting point for many students as they refined their scientific ideas.

RQ2: Relationship Between Student Idea Changes and KI

Regression of Change in KI on Changes in Ideas

We regressed the change in KI score on the change in total, accurate, vague and off task ideas about genetic inheritance and each time point across the revision activities (Figure 4). The timepoints included initial and revised responses to the within unit dialog; and initial, revised responses to the posttest dialog. For all time points, we observed similar regression coefficients. We found that the change in total number of ideas expressed was positively associated with a change in KI score (Appendix Table 4). The unstandardized regression coefficients of the change in the total number of ideas ranged between $b = .25$ and $b = .39$. This means that students who added one or more ideas achieved a progress in KI score of .25 or .39, on average (depending on the time points). The standardized regression coefficients ranged between $\beta = .40$ and $\beta = .55$, the determination coefficients R^2 ranged between $R^2 = .15$ to $R^2 = .30$; together this indicated a moderate to strong effect size. Thus, the role of the adaptive dialog in eliciting more student ideas, as described in RQ1, contributed to science learning.

To clarify the relationship between the changes in the number of ideas students expressed and KI progress, we examined changes in accurate, vague, and off-task ideas. The change in KI score was mainly driven by the change in the number of accurate ideas students integrated into their responses (Figure 4, Appendix Table 5). The standardized regression coefficients for accurate ideas ranged between $\beta = .45$ and $\beta = .65$ and the determination coefficients between $R^2 = .20$ and $R^2 = .43$ suggest a strong relationship between the adding of accurate ideas and KI change (Appendix Table 5). The change in the number of vague ideas was not associated with the change in KI Score at 7 of 8 time points and the change in number of off-task ideas was never associated with KI score (Appendix Tables 6, 7).

From a KI theoretical perspective, these results align with the view that students added accurate ideas to build on their prior knowledge and, often, to refine their vague ideas. Students initially expressed intuitive vague ideas such as ‘siblings share the same genes’ and used these ideas to discover and link to additional accurate ideas to elaborate their vague idea. The KI rubric rewarded students for integrating accurate ideas.

Idea Changes for Students Who Progress to Partial Versus Integrated KI

We identified the statistically significant idea changes for students who progressed to Partial KI (KI 3, 41% of students) and to Full KI (KI 4 or 5, 22% of students) in one of the revision timepoints (pretest-posttest, dialog) (Figure 5; Appendix Table 8). The remaining students did not progress or omitted responses. We illustrate the identified patterns qualitatively with examples.

Developing Partial KI: Adding and Partly Linking Ideas. Across the unit, students who progressed to Partial KI expressed an average of 1.2 ideas initially and 2.3 ideas in their revised responses. These students responded to the dialog by incorporating one or more ideas into their final response. Students were likely to add '*siblings inherit genes from their parents*' or to explain how variation in genetic inheritance impacts outcomes by adding '*alleles are combined differentially',* '*dominant versus recessive inheritance*' or '*punnett square*'. Further, students consistently dropped the vague idea that '*siblings look different because of vague, non-genetically inherited factors*' (Figure 5, Appendix Table 8).

Partial KI Example. We selected a representative example of a student who progressed to Partial KI from one of the participating schools with demographics reflective of the state. This student expressed new ideas in the dialog and partly linked these ideas in their revised response (Table 6).

This student began by expressing the idea that 'genes are inherited from the same parents' DNA' (rows 1,2). During the dialog, they shifted from their idea that siblings have the same DNA to the idea that because of their parents' different DNA, siblings will have different genes (row 4) and traits (row 6). They reiterated this view at the start of the posttest dialog, while also adding two new ideas with details about how parents' DNA affects the siblings' traits. They connected their initial idea about siblings' different DNA to a more detailed view involving different alleles, dominant/recessive inheritance, and phenotypes (rows 10, 12). They partially linked these ideas in their revised posttest response: "Because of their <u>parents' different alleles</u> they [siblings] have <u>different dominant and recessive traits</u> so it <u>changed them from looking exactly the same</u>."

Full KI Example: Distinguishing Between and Linking Ideas. Compared to students who progressed to Partial KI, students who progressed to Full KI started with more ideas and integrated more ideas. They expressed on average 2.1 ideas in their initial responses, 3.4 ideas during the dialog, and 4.75 ideas in their revised responses after the dialog. Thus, students who progressed to Full KI expressed double the number of ideas compared to those who progressed to Partial KI.

Students who progressed to Full KI significantly added ideas about the randomness of inheritance, as they refined their ideas about how siblings inherit genes (Figure 5, Appendix Table 8). They significantly added the vague idea that '*siblings share the same genes*' in the within unit dialog and elaborated this idea by significantly adding new ideas about the mechanisms of genetic variation such as *'chance in the genes each sibling inherits*', '*half of a sibling's genes are inherited from each parent*', and '*dominant/recessive genotypes'*. Many students also significantly added one or more ideas explaining why siblings display different traits: '*combinations of alleles*', '*punnett square'*, and '*genotype affects phenotype'*.

We selected a student from the Full KI subroup, using the same criteria as for the Partial KI group. This student progressed from vague to integrated knowledge (Table 7).

The student began at a KI 2 with two vague ideas:, 'same parents' and 'different genes' (rows 1-6). They clarified the link between these two ideas after the within-unit dialog by incorporating two mechanistic ideas regarding 'chance in what genes siblings inherit' and 'using a Punnett square to determine probability' (row 7). The student then distinguished between the two ideas in their initial view - that siblings have the same parents and different genotypes. In the posttest dialog they added new ideas about how inheritance results in display of different traits and that siblings inherit 'different combinations of alleles' (row 8), which can be 'dominant or recessive' from the parents' genes (rows 10, 12). After the posttest dialog, they linked these multiple ideas to distinguish between why siblings look different and also similar (row 13).

In summary, the analysis suggests that the students in the Partial KI group uniquely added the accurate idea '*siblings inherit genes from their parents*' more frequently than students in the Full KI group from pretest to posttest. They decreased their expression of the vague idea that siblings' differences are due to non-genetic factors. Distinguishing between these two ideas may represent an accessible starting point for accurate scientific reasoning. Students who progressed to Full KI added multiple and different accurate ideas regarding genetic variation (e.g.'*chance of inheriting different genes*', '*different allele combinations*'). Students used their accurate ideas to distinguish between their initial vague ideas that siblings have the same genes or different genes, elaborating how siblings' differently inherited genes from the same parents combine to dictate varied traits.

Student Perspectives on Dialog Design

Responding to a Student's Idea. Overall, students responded positively to the adaptive dialog. Eighty-one percent (22) of the students reported that they felt the thought buddy responded to *their* ideas. These students described how the dialog prompts encouraged them to examine their ideas about genetics. For example, one student reported, *"It asked me about my reasoning in a way more direct way…I had to really think about how do traits get really mixed up like that and oh that only happens sometimes…*" . When engaged in the dialog, students expressed that they recognized additional ideas they held and had not expressed initially. As one student reported, "*I didn't know the answer to the full question in the beginning but I learned it by the end [of the dialog]. The bot asked me different questions that made me think of material I learned in the beginning of this project that I had kind of forgotten.*" Another remarked, *"I used an idea that was kind of in the back of my mind but I had forgotten.*" Another shared a reflection process,*"...I tried to think of the [dialog] question to see if I had more ideas that I had not put in words yet.***".**

When asked ‘how did the dialog help you revise your response’ several of the students mentioned how they were continuing to consider their ideas and questions about genetic inheritance. As one student remarked, “*she helped me get ideas sometimes by talking about what siblings look like bc i have a sibling too…*”. Another commented, *“they got me thinking about it, why do siblings may have more of their dads genes, or moms genes or why some siblings just don’t look alike at all.*” One expressed wonders about differences between their siblings, “*I was born with curly hair and because of tradition we had to cut it off and now it has been thick and straight since 1.5 years, whereas my brother’s hair, it was straight from when he was born.*”

Talking to a more knowledgeable other. Of the 27 students, 44% (12) described talking to the thought buddy as similar to talking to a teacher or a more knowledgeable classmate. A student mentioned, *“It was almost the same as talking to a teacher, like the phrasing and questions seemed like the questions a teacher or a classmate would ask me in real life*.” Five of the twelve students distinguished the prompts in the adaptive dialog from the type of questions a peer would ask them. They reported that when talking with a peer they are often both stuck, or one student is looking to the other for ideas, whereas the adaptive dialog questions provoked them to think. One student noted, *“I guess these questions were more like what a professional would ask. Because it seems like it knows what it is doing. Where if I was talking to my friends, they would expect me to know the answer, but these seem like they are asking me to make me think more.”*

Design suggestions. Students who did not feel that the dialog responded to their ideas provided specific insights on how to improve the prompt design. 30% (8) of the students felt that talking to the thought buddy felt scripted and, or that the two prompts they got in the dialog were repetitive. In some cases, a first round prompt for one idea is similar to a second round prompt for another idea. A student described the dialog experience as scripted explaining, “*I feel like the teacher can understand you better. It's [the dialog] getting there but it would need a lot more* *work to be an actual teacher. Maybe you could incorporate ChatGPT to make its responses more human-like.*" Three students reported that the thought buddy ended the conversation too abruptly and requested that it continue the conversation. For example, students reported: "*It ended the conversation a little soon. I had more ideas to talk about..".* Or, *"It felt like I had it wrong when it ended the conversation…it would help if it asked more questions."*

In sum, the student interviews suggested that the adaptive dialog prompts generally affirmed the students' starting points in making sense of genetic inheritance. Students took the dialog seriously as a tool for considering their ideas. The dialog can be improved by ensuring that students experience the two prompts as different and by strengthening the sense of the dialog as a conversation with a beginning and end.

Study 2 Discussion

In this study, we analyzed how the NLP idea detection model and adaptive dialog, both guided by the KI framework, enabled students to integrate their ideas. The results of this study suggest that students are actively constructing knowledge in an NLP-based adaptive dialog by generating new, accurate disciplinary ideas and linking these ideas with their initial, often vague, ideas to improve their science explanations of genetic inheritance. Further, the adaptive guidance motivated students to add accurate ideas, in contrast to the Annotator study and other research that shows that students select new ideas through dialogic interactions but not necessarily more accurate ideas (Gerard & Linn, 2022; Chi & Wiley, 2014). These findings show the value of using KI to design NLP idea-detection dialogs, extending and clarifying related research on KI guidance design, teacher guidance, retrieval practice, and self-directed learning.

Guidance Design Tailored to Student Ideas

The findings show that the NLP idea-detection dialog can have positive effects for building integrated understanding in the ecologically valid context of middle and high school classrooms serving diverse students. Responding to prompts tailored to elicit further reasoning about the student's specific detected idea, resulted in students generating their own new ideas about genetic inheritance. Students then used some of the new ideas to elaborate their initial vague ideas or to distinguish between their ideas and construct links. Our study adds to the body of literature documenting the effectiveness of adaptive guidance for middle and high school science learning while also adding details about the process (Azevedo, Cromley et al., 2005; Gerard et al., 2015a; Vitale et al., 2016).

The logged NLP-based idea detection and KI scores provide empirical insights into students' KI processes. They show how students make sense of multiple ideas when learning, consistent with the significant increase in the number of ideas students expressed between their initial response and their responses to the dialog, and between their initial and final responses. This supports the view that learning involves generating and sorting out multiple ideas, rather than generating a single correct idea and eliminating incorrect ideas. When revising their responses students often elaborated their initially vague ideas rather than eliminating these ideas. This illustrates students' investment in their ideas and the benefits of guiding students to analyze and refine their ideas rather than offering alternatives that may interfere with productive reasoning.

The wide range of initial ideas students included in their explanations and the varied ideas they incorporated in response to adaptive prompts, provide empirical evidence for the variety of patterns of KI that occur in science classrooms. When prompted to build on their initial response students took advantage of distinct ideas to strengthen their understanding (Figure 5). Students who progressed to Full KI added one or more of 10 ideas, including significantly adding 6 different valid ideas to improve their explanation. Similarly, students who progressed to Partial KI added one or more of 12 ideas, including significantly adding one or more of 4 new ideas to integrate into their revised explanation. One idea was uniquely significantly added and one idea was uniquely significantly dropped by students who progressed to Partial KI. Three ideas were uniquely significantly added by students who progressed to Full KI. Thus, each student responded to the adaptive prompt by selecting the idea or ideas that fit their own explanation, not necessarily an idea that a classmate might choose. The many separate patterns of student idea integration aligns with the KI pedagogy that emphasizes supporting students to build on their own ideas.

Analysis of Partial and Full KI subgroups suggests that prompts for generating specific disciplinary ideas helped students to analyze the connection between their ideas and distinguish between initial contradictory or overlapping ideas. Many students (40-50%) who constructed integrated knowledge added the idea of '*chance in what you inherit*' or '*siblings receiving half of their genes from each parent*'. Students in the Partial KI subgroup expressed these ideas at a lower frequency than students in the Full KI subgroup (10%). Many students in the Full KI subgroup used one or both of these new ideas to distinguish among their initial, often inconsistent ideas. As shown in the examples, many expressed that '*siblings had the same genes*' and '*siblings had different genes*' in the same response. Using one of the new ideas about chance in inheritance, students clarified that siblings inherit genes from the same parents, and because of chance involved in which genes you inherit, siblings each inherit different allele combinations. Having made this distinction between their initial two vague ideas, students were then able to further elaborate how the inherited, distinct genotypes influenced siblings' traits.

Progress towards integrated knowledge occurred after one or two dialog experiences. Some students recognized one or both of the genetic variation ideas (chance of which genes you inherit; inherit half of your genes from each parent) during the within-unit dialog and others generated it during the post-test dialog. Yet, generating this specific idea about the role of randomness or chance was consistent for making an advance to integrated knowledge. Building on work on cognitive monitoring, which has shown to facilitate student learning with adaptive computer and human tutors (Greene & Azevedo, 2008), effectively monitoring one's conceptual understanding may involve analyzing the link or distinction between one's related ideas. Helping students to analyze related ideas to surface an idea they can use to distinguish between contrasting ideas is a possibility as we refine guidance for revision.

Overall, the dialog supported each student to follow the KI processes of eliciting ideas, discovering new ideas, distinguishing ideas, and sorting out ideas to create an explanation that connects their own ideas with evidence. These findings align with comments of effective teachers who remark on the unique reasoning paths their students take (Gerard et al, 2016). They also reinforce the value of using NLP to identify the distinct ideas each student holds.

Teacher Guidance

Consistent with the value placed on detecting student ideas that are grounded in their experiences and observations in the design of the NLP idea detection model and dialog guidance, students reflected in interviews that the dialog recognized *their* ideas and made them think about their perspectives. This aligns with the goals science teachers have for recognizing their students' ideas and providing guidance to help them build on their ideas (Luna, 2018). By identifying ideas students generated, the dialog led students to consider new ideas they held that they did not previously recognize as relevant. For example, one student said that they knew the ideas were "stuck in the back of my mind". These reflections on the value of the NLP idea-detection approach show that the dialog was personally relevant for the students. This finding aligns with research investigating how teachers formulate their guidance for students to be relevant to their students' insights (Luna, 2018).

Retrieval Practice

These results for NLP idea detection dialogs build on and suggest refinements to earlier research on retrieval practice, interleaving, and self-explanations. The dialog prompts students to retrieve information related to their detected idea and interleaves retrieval practice with additional activities, taking advantage of distributed practice and creating a desirable difficulty

(Birnbaum et al. 2013; Bjork & Bjork, 2020). The success of prompts asking students to refine their explanations by making connections between a detected idea and the material that has been studied, suggest ways to refine the benefits of self-explanations.

Tutoring and Self-Directed Learning

These findings for NLP idea detection dialogs extend prior research on tutoring, which found that adaptive scaffolding was more effective than fixed scaffolding because in large part it activated students' prior knowledge (Azevedo et al., 2005). In this study, when analyzing their own ideas, students significantly added new ideas in the dialog and these new ideas were most often accurate. Students followed multiple paths to generate coherent responses, as supported by the significant increase in the frequency of expression of all but one of the accurate detectable ideas in the revised responses.

Developing self-directed learners requires changing student expectations as well as practice (Zimmerman, 2013). The Annotator (Gerard & Linn, 2022) and the dialog both provided students a first step towards engaging in self-directed learning. Nevertheless, on the posttest in this study, students still benefited from the dialog to improve their responses rather than spontaneously engaging in self-directed learning. Only a few students gave an initial response when prompted by the posttest dialog that incorporated all the ideas they had generated in their revised response to the within unit dialog. These terse responses may reflect student experiences with typical classroom tests that ask for recall of details rather than requiring the integration of multiple ideas.

Limitations of Study 1 and Study 2

The findings from these studies apply mainly to populations that are very similar to the students and teachers involved. More evidence is needed to extend these findings to new populations.

A possible limitation of using NLP to promote knowledge integration is that some student ideas may not be detected. In Study 1, four categories of student ideas that appeared sparsely in the training data were dropped from the NLP idea-detection model. Further, we recognize that there will always be new, student ideas that were not a part of the training data and hence the NLP model is not trained to detect. In Study 2, if a student expressed an undetectable idea in the dialog they received a generic guidance prompt designed to promote knowledge integration, rather than an adaptive prompt guiding the student to build on their specific idea. The generic prompt encouraged reflection and revisiting evidence in the unit. We are exploring the design of adaptive guidance for situations when the NLP model cannot detect an idea by taking advantage of what the student may have responded to previous questions. We are working on how to update the idea-detection models rapidly using generative AI techniques to incorporate new student ideas we gather in partnership with teachers.

An open question concerns the impact of each adaptive guidance prompt impacting knowledge integration trajectories. We studied the overall impact of the prompts in the dialog. A fine grained analysis of the effectiveness of each prompt requires a larger data set, in order to group students who received the same prompt and track progress at the idea-level.

Practical Implications

Engaging students in an adaptive dialog in a web-based science unit, offers teachers a promising instructional resource for encouraging all their students to integrate their ideas. The accuracy of the NLP idea detection in new school contexts, coupled with the evidence of student learning gains and students' reported experiences with the dialog, suggests that it is possible to scale support for knowledge integration by using NLP tools embedded in free, open-source curriculum. The dialog supports teachers to provide each of their 30-35 students across 5-7 class periods totaling 150 students per teacher, personalized prompts to reflect on their initial ideas about a standards-aligned science concept, discover new ideas they hold, and link these ideas to refine their understanding.

By partnering with teachers to design and refine the NLP idea detection rubrics and prompts for their specific students the adaptive dialogs detected and responded to ideas that the teachers value. Teacher input to the NLP idea detection and prompt design contributed to the students' feeling like their ideas were recognized by the avatar and that they could respond. As a result, students reported feeling encouraged to value and refine their initial ideas.

Conclusions

In conclusion, AI enhanced adaptive dialogs offer promise as a mechanism for strengthening students' learning opportunities to both build scientific knowledge and the self-directed practice of revising their explanations. Further research is needed to investigate and improve the dialog. Capitalizing on the unit and revision activities that promoted distinguishing between ideas would be valuable. This might involve designing hints that help students distinguish and integrate their ideas. We might extend the NLP method to individually guide Partial KI and Full KI students. This work validates the use of the knowledge integration pedagogical framework to design guidance, analyze student learning, and generate new research questions.

References

Alhindi, T., & Ghosh, D. (2021). "Sharks are not the threat humans are": Argument Component Segmentation in School Student Essays. *arXiv preprint arXiv:2103.04518*

Atteberry, A., Loeb, S., & Wyckoff, J. (2017). Teacher Churning: Reassignment Rates and Implications for Student Achievement. *Educational Evaluation and Policy Analysis*, *39*(1), 3-30

Azevedo, R., Cromley, J. G., Winters, F. I., Moos, D. C., & Greene, J. A. (2005). Adaptive human scaffolding facilitates adolescents' self-regulated learning with hypermedia. *Instructional Science, 33*(5-6), 381–412.

Azevedo, R., & Cromley, J. G. (2004). Does Training on Self-Regulated Learning Facilitate Students' Learning With Hypermedia? *Journal of Educational Psychology, 96*(3), 523–535. https://doi.org/10.1037/0022-0663.96.3.523

Azevedo, R., Moos, D., Johnson. A., & Chauncey, A (2010) Measuring Cognitive and Metacognitive Regulatory Processes During Hypermedia Learning: Issues and Challenges, *Educational Psychologist*, 45:4, 210-223, DOI: 10.1080/00461520.2010.515934

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

Banilower, E. (2019). Understanding the big picture for science teacher education: The 2018 NSSME#. *Journal of Science Teacher Education, 30*(3), 201–208.

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*, 1082–1112.

Bexte, M., Horbach, A., Torsten, Z., (2022). Similarity-Based Content Scoring - How to Make S-BERT Keep Up With BERT. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, 118–123, Seattle, Washington. Association for Computational Linguistics.

Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied research in Memory and Cognition*, 9(4), 475

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & cognition, 41*, 392–402.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education, 5(1), 7–74.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*, 245–281. http://dx.doi.org/10.3102/00346543065003245

Butler, A. C., & Roediger, H. L., III. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*, 604–616. http://dx.doi.org/10.3758/MC.36.3.604

Campbell, T., Schwarz, C., & Windschitl, M. (2016). What we call misconceptions may be necessary stepping-stones toward making sense of the world. *Science and Children, 53*(7), 69–74.

Clark, D. B. (2006). Longitudinal conceptual change in students' understanding of thermal equilibrium: An examination of the process of conceptual restructuring. *Cognition and Instruction, 24*(4), 467–563.

diSessa, A. A., & Minstrell, J. (1998). Cultivating conceptual change with benchmark lessons. In J. G. Greeno & S. Goldman (Eds.), *Thinking practices* (pp. 155-187). Lawrence Erlbaum Associates.

Chi, M. T. H. (2009). Active-Constructive-Interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, *1*, 73–105.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145–182.

Chi, M. T., De Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive science, 18*(3), 439–477.

Chi, M. T., Kang, S., & Yaghmourian, D. L. (2017). Why students learn more from dialogue-than monologue-videos: Analyses of peer interactions. *Journal of the Learning Sciences, 26*(1), 10-50.

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*, 219–243.

Clark, D. B. (2006). Longitudinal conceptual change in students' understanding of thermal equilibrium: An examination of the process of conceptual restructuring. *Cognition and Instruction, 24(4)*, 467-563.

Cogliano, M., Bernacki, M. L., & Kardash, C. M. (2021). A metacognitive retrieval practice intervention to improve undergraduates' monitoring and control processes and use of performance feedback for classroom learning. *Journal of Educational Psychology, 113*(7), 1421–1440. https://doi.org/10.1037/edu0000624

Coll, R. K., France, B., & Taylor, I. (2005). The role of models/and analogies in science education: Implications from research. *International Journal of Science Education, 27*(2), 183–198.

Crawford, L., Lloyd, S., & Knoth, K. (2008). Analysis of student revisions on a state writing test. *Assessment for Effective Intervention, 33*, 108–119.

Davis, E. (2003). Prompting middle school science students for productive reflection: Generic and directed prompts. *The Journal of the Learning Sciences, 12*(1), 91–142.

Davis, E. A., & Krajcik, J. S. (2005). Designing Educative Curriculum Materials to Promote Teacher Learning. *Educational Researcher*, *34*(3), 3–14

Firth, J., Rivers, I., & Boyle, J. (2021). A systematic review of interleaving as a concept learning strategy. *Review of Education*, 9(2), 642-684.

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.

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

Gerard, L. F., & Linn, M. C. (2016). Using automated scores of student essays to support teacher guidance in classroom inquiry. Journal of Science Teacher Education, 27(1), 111-129. doi:10.1007/s10972-016-9455-6

Gerard, L., Matuk, C., McElhaney, K., Linn, M.C. (2015a). Automated, adaptive guidance for K-12 education. *Educational Research Review, 15*, 41–58.

Gerard, L., Ryoo, K., McElhaney, K., Liu, L., Rafferty, A., & Linn, M. C. (2015b). Automated guidance for student inquiry. *Journal of Educational Psychology, 108*(1), 60–81.

Gerard, L., Wiley, K., Debarger, A.H., Bichler, S., Bradford, A., & Linn, M.C. (2022). Self-directed Science Learning During COVID-19 and Beyond. J Sci Educ Technol 31, 258–271.

Greene, R. & Azevedo, R. (2010) The Measurement of Learners' Self-Regulated Cognitive and Metacognitive Processes While Using Computer-Based Learning Environments. *Educational Psychologist, 45*(4), 203–209. DOI: 10.1080/00461520.2010.515935

Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., & Khot, T. (2023). "Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs." ArXiv arXiv:2311.04892v1

Harrison, E., & Gerard, L. (2018). Encouraging revision of scientific ideas with critique in an online genetics unit. In Kay, J. & Luckin, R. (Eds.), *Rethinking Learning in the Digital Age: Making the Learning Sciences Count, 13th International Conference of the*

Learning Sciences (ICLS) 2018, Volume 1, pp. 816-824. London, UK: International Society of the Learning Sciences

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112. http://dx.doi.org/10.3102/003465430298487

Horbach, A., & Zesch, T. (2019). The Influence of Variance in Learner Answers on Automatic Content Scoring. Frontiers in Education, 4, https://doi.org/10.3389/feduc.2019.00028

Johnson, J.M., Khoshgoftaar, T.M. (2019). Survey on deep learning with class imbalance. *J Big Data* 6, 27.

Kali, Y. (2006). Collaborative knowledge building using the Design Principles Database. *Computer Supported Learning* 1, 187–201

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254–284. http://dx.doi.org/10.1037/0033-2909.119.2.254

Koedinger, K.R., Aleven, V. (2007). Exploring the Assistance Dilemma in Experiments with Cognitive Tutors. *Educ Psychol Rev* 19, 239–264. https://doi.org/10.1007/s10648-007-9049-

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Merrill, J., Urban-Lurain, M., Doherty, J. H., & Haudek, K. C. (2020). Comparison of Machine Learning Performance Using Analytic and Holistic Coding Approaches Across Constructed Response Assessments Aligned to a Science Learning Progression. *Journal of Science Education and Technology*, 1–18.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*, 1–54.

Kyza, E. (2009). Middle-School students' reasoning about alternative hypotheses in a scaffolded, software-based inquiry investigation. *Cognition and Instruction, 27*(4), 277–311.

Lee, HS., Gweon, GH., Lord, T. *et al.* (2021) Machine Learning-Enabled Automated Feedback: Supporting Students' Revision of Scientific Arguments Based on Data Drawn from Simulation. *J Sci Educ Technol* 30, 168–192. https://doi.org/10.1007/s10956-020-09889-7

Lee, D.-H., Pujara, J., Sewak, M., White, R. W., & Jauhar, S. K. (2023). Making Large Language Models Better Data Creators. *EMNLP*. *arXiv preprint arXiv:2310.20111*

Linn, M.C., Donnelly-Hermosillo, D. & Gerard, L. (2023). Synergies between learning technologies and learning sciences: Promoting Equitable Secondary Science Education. Handbook of Research on Science Education Volume III, Routledge Press

Linn, M., & Eylon, B.-S. (2011). Science learning and instruction: Taking advantage of technology to promote knowledge integration. New York: 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. https://doi.org/10.1080/10627190801968224

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.

Liu, L., Ryoo, K., Linn, M.C., Sato, E. & Svihla, V. (2015). Measuring Knowledge Integration Learning of Energy Topics: A two-year longitudinal study. *International Journal of Science Education, 37*(7), 1044–1066. DOI: 10.1080/09500693.2015.1016470

Luna, M. (2018) What Does it Mean to Notice my Students' Ideas in Science Today?: An Investigation of Elementary Teachers' Practice of Noticing their Students' Thinking in Science. *Cognition and Instruction, 36*(4), 297–329, DOI: 10.1080/07370008.2018.1496919

Matuk, C., Linn, M.C. Students' perceptions of the impacts of peer ideas in inquiry learning. *Instr Sci* 51, 65–102 (2023). https://doi.org/10.1007/s11251-022-09607-3

Matuk, C., Zhang, J., Uk, I., & Linn, MC. (2019). Qualitative graphing in an authentic inquiry context: How construction and critique help middle school students to reason about cancer. *J Res Sci Teach*, *56*, 905–936.

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

Morales-Doyle, D. (2017). Justice-centered science pedagogy: A catalyst for academic achievement and social transformation. *Science Education, 101*(6), 1034–1060.National Center for Education Statistics (2018). National teacher and principal survey. Retrieved from: https://nces.ed.gov/surveys/ntps/tables/ntps1718_fltable06_t1s.asp

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology*, 21(1), 183–196.

Nickerson, R. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220. https://doi.org/10.1037/1089- 2680.2.2.175

Obaid T., Aghajani H., Linn M.C. (2023). Using optimized clustering to identify students' science learning paths to knowledge integration. *STEM Edu Rev*iew.

Osborne, J. (2014). Scientific practices and inquiry in the science classroom. In N. G. Lederman, & S. K. Abell (Eds.), *Handbook of research on science education* (Vol. II, pp. 579–599). New York: Routledge.

Pi, Z., Zhang, Y., Shi, D., Guo, X., & Yang, J. (2022). Is self‐explanation better than explaining to a fictitious student when learning from video lectures?. *British Journal of Educational Technology*, 53(6), 2012-2028.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33.

Puntambekar, S., Stylianou, A., & Goldstein, J. (2007). Comparing classroom enactments of an inquiry curriculum: Lessons learned from two teachers. *Journal of the Learning Sciences, 16*(1), 81–130. doi:10.1080/10508400709336943

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. R version 4.2.1. https://www.R-project.org/.

Rafferty, A. N., Gerard, L., McElhaney, K., & Linn, M. C. (2013). Automating Guidance forStudents' Chemistry Drawings. Proceedings of the Workshops at the 16th International Conference on Artificial Intelligence in Education AIED 2013 (pp. 1-8).

Resnick, L. B., Asterhan, C. S., & Clarke, S. N. (2018). Accountable talk: Instructional dialogue that builds the mind. Geneva, Switzerland: *The International Academy of Education (IAE) and the International Bureau of Education (IBE) of the United Nations Educational*, Scientific and Cultural Organization (UNESCO).

Richland, L. E., Bjork, R. A., Finley, J.R., Linn, M.C. (2005). Linking Cognitive Science to Education: Generation and Interleaving Effects. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds). *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (pp. 1850-1855). Mahwah, NJ: Lawrence Erlbaum Associates.

Richland, L. E., Linn, M. C., & Bjork, R. A. (2007). Cognition and instruction: Bridging laboratory and classroom settings. In F.T. Durso (Ed.), *Handbook of Applied Cognition*, Second Edition (pp. 555-583). West Sussex, England: John Wiley & Sons, Ltd.

Riordan, B., Bichler, S., Bradford, A., King Chen, J., Wiley, K., Gerard, L., & Linn, M. C. (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,* Association for Computational Linguistics, 135–144.

Rivard, L. P. (1994). A review of writing to learn in science: Implications for practice and research. *Journal of Research in Science Teaching, 31*(9), 969–983.

Roediger, H. & Karpicke, J. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention, *Psychological Science, 17*(3).

Rosebery, A. S., Ogonowski, M., DiSchino, M., & Warren, B. (2010). “The coat traps all your body heat”: Heterogeneity as fundamental to learning. *The Journal of the Learning Sciences, 19*(3), 322–357.

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers’ informal formative assessment practices and students’ understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 44*(1), 57–84.

Ryoo, K., & Linn, M. C. (2014). Designing guidance for interpreting dynamic visualizations: Generating versus reading explanations. *Journal of Research in Science Teaching, 51*(2), 147–174. https://doi.org/10.1002/tea.21128

Ryoo, K., & Linn, M. C. (2016). Designing automated guidance for concept diagrams in inquiry instruction. *Journal of Research in Science Teaching, 53*(7), 1003–1035.

Schulz, C., Meyer, C., Sailer, M., Kiesewetter, J., Bauer, E., Fischer, F., Fischer, M., & Gurevych, I. (2019). Challenges in the Automatic Analysis of Students’ Diagnostic Reasoning. In Proceedings of AAAI

Smith, J. P., diSessa, A. A., & Roschelle, J. (1993-1994). 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

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., et al. (2019). Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education, 131*, 33–48. https://doi.org/10.1016/j.compedu.2018.12.005

Tansomboon, C., Gerard, L., Vitale, J., & Linn, M. C. (2017). Designing automated guidance to promote productive revision of science explanations. *International Journal of Artificial Intelligence in Education (IJAIED), 27*, 1–29.

Tunstall, L., von Werra, L., & Wolf, T. (2022). *Natural Language Processing with Transformers*. Sebastopol, CA: O'Reilly Media, Inc.

Vakil, S., & Higgs, J. (2019). It's about power. *Communications of the ACM*, *62*(3), 31-33.

VanLehn, K. (2011). The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist, 46*(4), 197-221.

Vitale, J., 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, DOI: 10.1080/09500693.2016.1198969

Williams, M. (2008). Moving Technology to the Center of Instruction: How One Experienced Teacher Incorporates a Web-based Environment Over Time. *J Sci Educ Techno*l 17, 316–333.

Williams, M., DeBarger, A. H., Montgomery, B. L., Zhou, X., & Tate, E. (2012). Exploring middle school students' conceptions of the relationship between genetic inheritance and cell division. *Science Education, 96*(1), 78–103.

Zhai, X, C. Haudek, K, Shi, L, H. Nehm, R, Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *J Res Sci Teach*, *57*, 1430–1459. https://doi.org/10.1002/tea.21658

Zhai, X., He, P., & Krajcik, J. (2022). Applying machine learning to automatically assess scientific models. *Journal of Research in Science Teaching*, 59(10), 1765–1794

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.

Zimmerman, B. (2013) From Cognitive Modeling to Self-Regulation: A Social Cognitive Career Path, *Educational Psychologist*, 48:3, 135-147, DOI: 10.1080/00461520.2013.794676

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*. https://doi.org/10.1016/j.compedu.2019.103668

Tables and Figures

Table 1

Knowledge Integration Rubric for: "Why Do Siblings From the Same Parents Look Similar But Not Exactly the Same?"

KI	Description	Examples
1	**Off task**	I don't know[4]; dgddgdgd [typing of random letters]
2	**Vague** Non-normative ideas; Or vague ideas that cannot discern if accurate or inaccurate	Because you and your sibling have close genes but they are not the same You inherit similar amounts of the same traits from the same parents at slightly different amounts. "I think it could be because they are conceived at two different times, so the DNA they got was different.
3	**Partial link** A normative idea linked to vague idea(s) or isolated normative idea	[vague and normative idea linked] Since siblings have the same parents they have the same chances to end up with certain genes. For example it could be a 25% chance of the child getting blue eyes and only one of the children ends up with blue eyes and it could be a 50% chance the child ends up with brown hair then both end up with it. that's why they look similar because they have the same chances for the same genes. (same parents=vague; chance of inheriting genotype=accurate) [isolated normative] siblings look similar to each other in some ways because they both get genetics from the mom and the dad but they dont look the same.
4	**Complete** One full link between two normative scientific ideas	"I think they look the same but not exactly the same because they got Rr and the other sibling would get a rr so that would mean that they have the same parents but different genes." Siblings look similar in some ways but they do not look the same because their genes come from the same people but they do not have the same exact genotypes. For example, if a parent's genotypes for dimples are Dd and dd the children's likely hood of having dimples would be about 50%.
5	**Complex** Links among three or more normative scientific ideas	"Siblings look similar because they get most of the same traits from their parents but you have a fifty percent chance of getting a dominant trait if it is big R, little r and little r, little r so one sibling could get dominant one get recessive. The parents can have different combinations of alleles for different traits. For example, eyes, the mom can be heterozygous with one dominant allele and one recessive allele which the dominant allele masks the looks of the recessive. This cause the dominant allele to only show

[4] We cannot distinguish between "I don't know" as an indicator of students' monitoring their comprehension versus disengaging. We have grouped IDK with other uninterpretable responses.

Table 2

Sample of the Detectable Ideas, Reflective Student Responses, and Dialog Prompts

ID	Detectable Idea	Sample Student Responses (Idea Detected Underlined)	Guidance Prompts [Prompt 2 assigned if idea detected twice]
Sample Vague Ideas			
3	Siblings have different DNA /Genes / Chromosomes	I believe that siblings look similar to each other because they have some similar traits from each parent, <u>but not the exact same DNA code</u>. I think siblings look similar to each other but not the same because <u>there are so many different ways DNA can be combined</u>, so no one will look exactly the same as their sibling.	1: In what ways do siblings have different genetic information? 2: Why do siblings look more alike than people who are not related?
4	Vague differences (ex: time born, gender)	Because they can get different alleles and environmental pressures, but their dna is still similar. <u>also, they could be different ages or gender</u>. they look similar because they have anelle genes from each parent <u>and not born at the same time so they all have different affects</u> on them.	1: One sister has dimples and the other does not. Why/How do you think this happened? 2: Why might one sister look more like their mother and the other sister look more like their father?
6	Siblings have the same parents	because they <u>formed in the same amniotic sac</u>. the <u>genotypes in the same parents go to the kids</u> which makes them similar.	1: Why does it matter that siblings have the same parents? 2: Why does it matter that the parents are the same?
Sample Accurate Ideas			
16	Chance / Randomness	They inherit different traits from their parents <u>because sometimes there can be a 50% chance a child will inherit a trait from a parent</u>. Because <u>of the half genes (from the parents) that gets passed down are random</u>. This means that the siblings don't necessary have to inherit the same traits.	1: How are the genes you inherit randomly selected? 2: How might you predict what traits a child might have?
18	Everyone has different combinations of alleles	<u>there are so many different ways DNA can be combined</u>, so no one will look exactly the same as their sibling. Even though they have the same parents, the parents might give the kids different genotypes for different traits. the <u>genes get mixed</u>	1: How does the combination of alleles that they get from each parent determine what they look like? 2: How could you predict what combination of alleles a sibling would get from their parents?

Table 3

Ideas Detected and Evaluation of Model With Training Data, Organized by Category and Frequency on Pretest (lowest to highest)

		Evaluation Metrics			Label Count	
Idea	Idea Descriptor	F-score	Precision	Recall	human	NLP
Vague Ideas						
5. Unique	Everyone is unique/different	0.541	0.619	0.481	31	26
7. Uneven inheritance	Uneven inheritance/more genes from one parent	0.874	0.852	0.897	174	183
4. Gender/Age	Vague differences (Age, Gender)	0.6991	0.746	0.658	76	67
15. Different traits from parents	Different traits/phenotype come from each parent	0.623	0.620	0.626	1731	108
9. Genes similar	Genes/dna/traits are similar but not exactly the same	0.713	0.713	0.713	94	94
6. Same parents	Siblings have the same parents	0.874	0.852	0.897	174	183
8. Same genes	Siblings share the same genes/dna/chromosomes	0.786	0.786	0.786	257	257
3. Different genes	Siblings have different DNA/Genes/Chromosomes	0.770	0.754	0.786	485	505
Off-Task						
2. Repeats question	Repeats the question without adding a response	0.281	0.308	0.258	31	26
Accurate ideas						
19. Difft Phenotype	Siblings have different phenotypes	0.396	0.528	0.528	60	36
24. 50/50 inherit	Half genes from each parent	0.897	0.897	0.897	117	117
25. Punnett	Describe how a Punnett Square works	0.657	0.657	0.657	35	35
23. Geno/Phenotype	Genotype affects phenotype	0.565	0.547	0.583	108	115
16. Chance	Chance/Randomness of inherited genes	0.825	0.793	0.859	85	92
17. Dom/Recessive	Dominant vs. recessive genes	0.877	0.877	0.877	65	65
18. Allele combos	Everyone has different combinations of alleles	0.708	0.672	0.749	243	271

20. Genes from parents	Genes/DNA/traits are inherited from parents	0.881	0.870	0.893	599	615

Table 4

Study Participants and Demographics for the Five Schools

School	No. of Partner Teachers	No. of Participating Students	Grade	School-level Demographics		
				Free Lunch	ELL	Non-White
1	1	51	9th	95%	33%	99%
2	2	75	9th	26%	5%	66%
3	2	121	8th	62%	33%	80%
4	2	93	8th	22%	11%	67%
5	2	270	6th	13%	3%	50%

Table 5

KI Score Mean, Median, and Standard Deviation

Timepoint	*Mean*	*SD*	*Median*
Pretest to posttest			
Pretest explanation	2.84	0.78	3.00
Revised posttest response [a]	3.06	0.79	3.00
Within-unit dialog			
Initial response	2.66	0.62	3.00
Revised response [a]	3.04	0.78	3.00
Posttest dialog			
Initial response	2.62	0.68	3.00
Revised response [a]	3.06	0.79	3.00

Note. [a] Significant change from previous time point.

Table 6

Example of a Student Who Progressed to Partial KI

Speaker	Dialog Responses	Ideas Detected
Pretest		
Student	because of dna, they have similar dna in each other	Same genes [8]; KI 2
Within-unit dialog		
Student initial response	because of their parents DNa	Genes inherited from parents [20]
Thought buddy	*How do the genes children get from their parents determine which traits they'll have?*	
Student	by theire genotypes	Different genes [3]
Thought buddy	*In what ways are siblings' genes different?*	
Student	diffrent size and color	Different phenotypes [19]
Student revised response	its because of their parents different dna so they will have diffrent size or body size or hair color	Nothing detected - 'genes inherited from parents' evident; KI 2
Posttest dialog		
Student initial response	Because of their parents different Dna	Nothing detected
Thought buddy	*Let's think about how this happens together! How do you think the parents have an affect on what the two siblings look like?*	
Student	I think it has an affect by their dominant and recessive parts	Dominant / Recessive [17]
Thought buddy	*How can the combination of alleles make one sibling look different from the other?*	
Student	diffrent alleles	Combinations of alleles [18]
Student revised response	Because of their parents different alleles they have diffrent dom and recessive traits so it changed them from looking exactly the same	Dominant / Recessive [17]; Different combinations of alleles [18]; Genotype affects phenotype [23]; KI 3

Table 7

Example of a Student Who Progressed to Full KI

Speaker	Dialog Responses	Ideas Detected
Pretest		
Student	Because they have the same parents but not the exact same genetics.	Different genes [3]; Same parents [6]; KI 2
Within-unit Dialog		
Student initial response	Siblings look similar in some ways but they do not look the same because they do not have the same of their genotypes.	Different genes [3]; KI 3
Thought buddy	*In what ways are siblings' genes different?*	
Student	They have different genotypes.	Different genes [3]
Thought buddy	*Why do siblings look more alike than people who are not related?*	
Student	Because their genes come from the same people.	Same parents [6]
Student revised response	Siblings look similar in some ways but they do not look the same because their genes come from the same people *but* they do not have the same exact genotypes. For example, if a parent's genotypes for dimples are Dd and dd the children's likely hood of having dimples would be about 50%.	Different genes [3]; Same parents [6]; Chance [16]; Punnett Square [25]; KI 4
Posttest Dialog		
Student initial response	Siblings look different from each other because they inherent different combinations of alleles from their mother and father.	Different combinations of alleles [18]; Genes inherited from parents [20]; KI 3
Thought buddy	*How do the genes children get from their parents determine which traits they'll have?*	
Student	It really depends on whether the genes passed down are dominant or recessive. For example if both parents have brown eyes their offspring have really high chances of having brown eyes. Works the same way for other traits.	Chance [16]; Dominant / Recessive [17]
Thought buddy	*How can the combination of alleles make one sibling look different from the other?*	
Student	If one parent has recessive alleles for a certain trait but the other parent has heterozygous dominant alleles for the same trait it would give the offspring a higher percentage of having the recessive trait however there is still a chance to have the dominant trait.	Chance [16]; Dominant / Recessive [17]

Speaker	Dialog Responses	Ideas Detected
Student revised response	The reason siblings look different from each other is they inherit different combinations of alleles from their mother and father. For example, if one parent has recessive alleles for a certain trait but the other parent has heterozygous dominant alleles for the same trait it would give the offspring a higher percentage of having the recessive trait however there is still a chance to have the dominant trait. The many combinations make it possible for siblings to look similar but still have some traits that differentiate them from each other.	Chance [16]; Dominant/Recessive [17]; Different combinations of alleles [18]; Genes inherited from parents [20]; Genotype affects phenotype [23]; KI 4

Figure 1

Siblings Prompt and Sample NLP-Based Adaptive Dialog

You may notice that biological siblings look similar, but they don’t always look exactly the same (unless you are an identical twin!). For example these biological siblings look similar, but not exactly the same. *Why do you think siblings look similar to each other in some ways, but not exactly the same?*

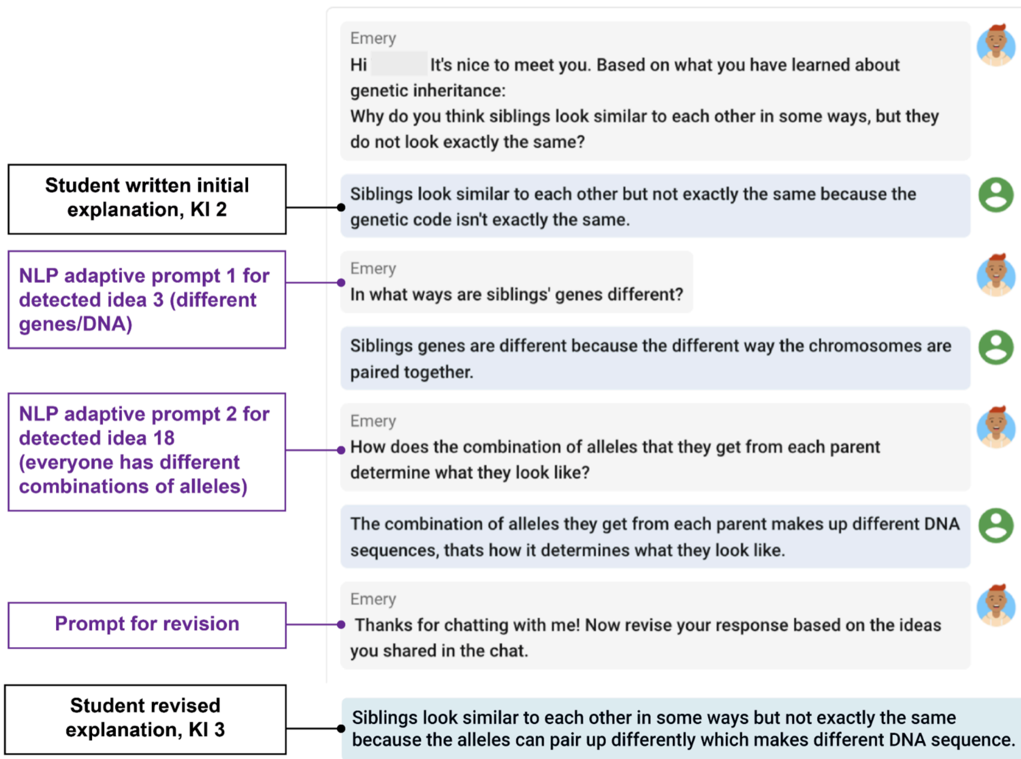


Figure 2

Timeline of Data Collection

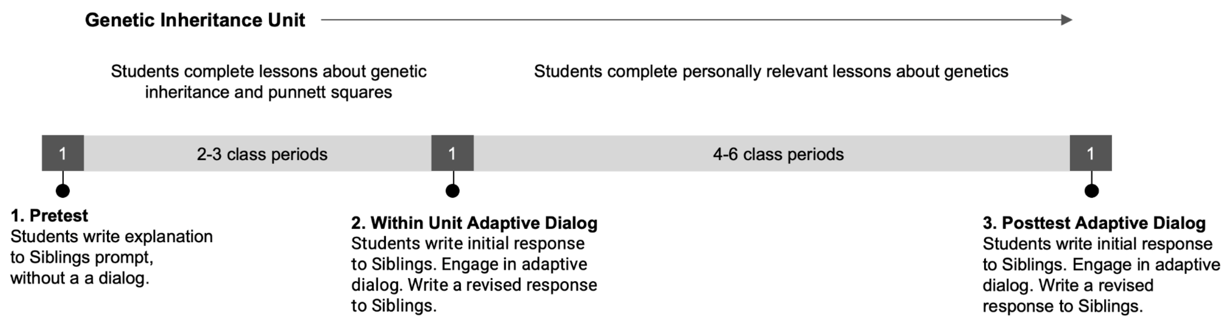
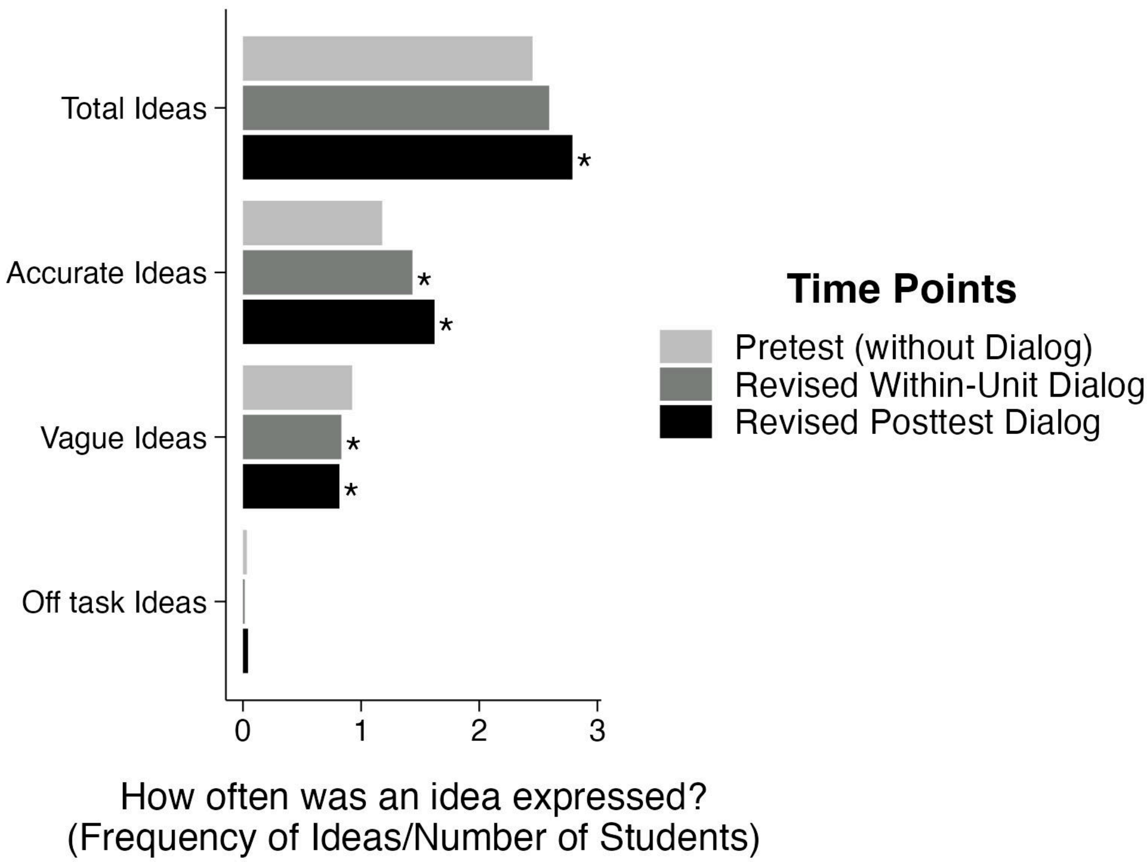


Figure 3

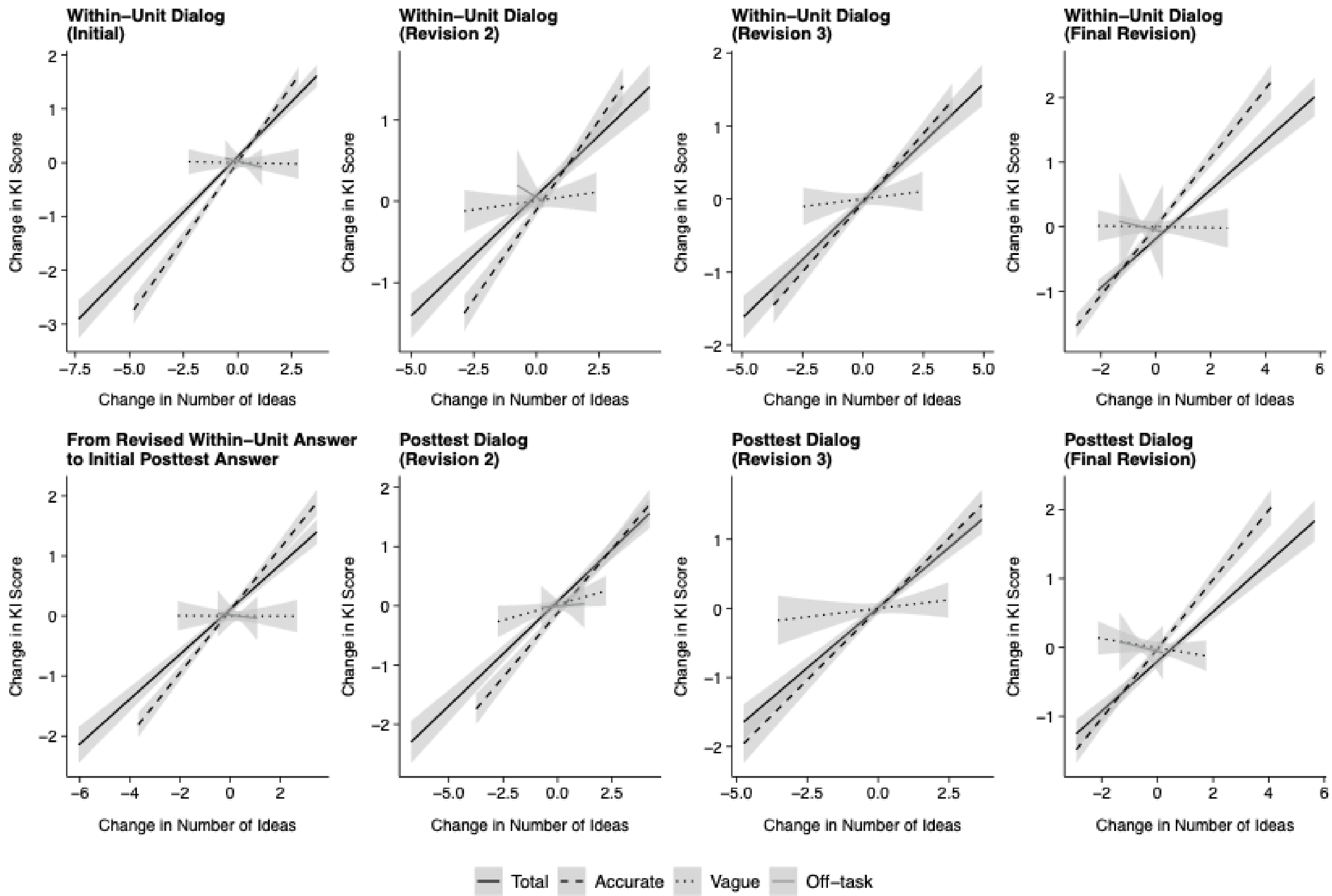
Frequency of Student Ideas Expressed at Each Timepoint



Note. Due to the low frequency, off task ideas were not tested for significance.

Figure 4

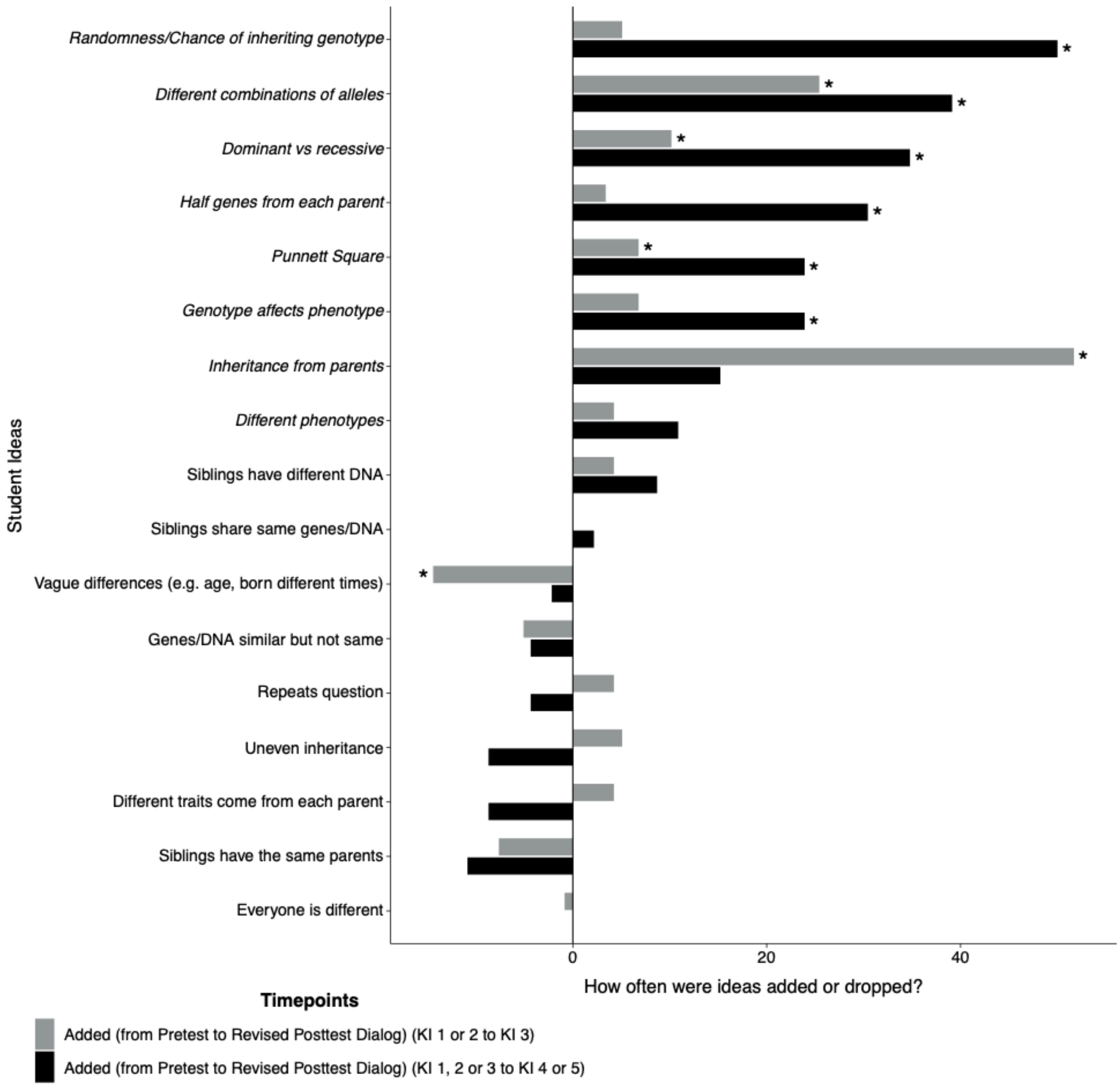
Regression of Change in KI on Changes in Ideas (Standardized).



Note: Shaded area around each line represents the standard error (SE).

Figure 5

Significant Changes of Ideas for Students who Made Different Progress in KI Scores: Pretest and Revised Response in Posttest Dialog



Note. N (from KI 1 or 2 (Pretest) to KI 3 (Revised Posttest Answer)) = 118, *N* (from KI 1, 2 or 3 (Prestest) to KI 4 or 5 (Revised Posttest Answer)) = 46; Bars with * indicate that the idea was significantly added or dropped in this subgroup. Accurate ideas are italicized.

Appendix

Additional NLP Modeling Details

General. Deep learning models for NLP applications are typically based on pre-trained transformer models. Transformer models typically employ word representations that are learned from language models trained on large text corpora. A language model learns to predict words in the input corpus which are either masked out or constitute the next word in the input during training. Such “self-training,” when paired with the transformer neural network architecture, produces word representations that are useful across many NLP tasks, including text classification and information extraction. Transformer networks use “self-attention” between word-like representations that take into account the context of words' usage.

Holistic scoring model details. Student responses were tokenized with the spaCy English tokenizer. The basic machine learning model architecture was: a pre-trained transformer model embedded a response’s tokens (using the “bert-base-uncased” pre-trained instance) (Tunstall et al., 2022); a single vector representing the response was generated (using the output vector for the "CLS" special token); dropout was applied (0.1); a single linear layer with sigmoid activation generated an unnormalized scalar logit.

Holistic scoring model training. During training, response scores were scaled to [0, 1]; for evaluation, the scaled scores were converted back to their original range. The model’s training objective was minimization of a mean squared error loss function.

Idea detection model details. Student responses were tokenized with the spaCy English tokenizer. For idea detection, a multi-label sequence labeling model was employed: response tokens were embedded with the same base language model as in holistic scoring; the response representation was passed through a simple sequence-to-sequence recurrent neural network (RNN; 1-layer Gated Recurrent Unit) to further contextualize token; for each token, a linear layer with sigmoid activation produced an unnormalized scalar logit for each idea category.

Idea detection model training. The network was trained on per-token idea category labels (see Section “Machine learning model development”). The model’s training objective was minimizing a sequence-based cross-entropy loss.

General model training. Models were trained with 10-fold cross validation with train/validation/test splits (80%/10%/10%, respectively). Demographic data was not available at training time for data splitting or stratification. Predictions were pooled (concatenated) across folds and used for evaluation. The models’ hyperparameters were tuned as follows: models were trained on each train split and evaluated on the validation split, keeping the best predictions across epochs; predictions were pooled from all folds on the validation sets, performance was evaluated, and the best-performing configuration of hyperparameters was selected. Following hyperparameter tuning, full models were trained on combined train and validation splits with 10-fold cross-validation to the median best epoch across folds from the hyperparameter tuning phase. Full model performance was evaluated on the pooled predictions from the test splits. This training and evaluation procedure has two main advantages: 1) increasing the stability of estimates of performance during both the hyperparameter tuning and full model testing phases; 2) using more data for training and evaluating the final models in order to provide better estimates of model performance.

Both model types (holistic scoring and idea detection) were optimized with the Adam optimizer, a learning rate tuned from the set {2e-5, 3e-5, 5e-5}, and a batch size 16. Hyperparameters were tuned for 20 epochs. An exponential moving average was used to smooth the model's weights across training epochs (decay rate = 0.999)

Table 1

Evaluation of Classroom Data F-1 scores

		Evaluation Metrics			Label Count	
Idea	Idea Descriptor	F-score	Precision	Recall	human	NLP
Vague Ideas						
5. Unique	Everyone is unique/different	1	1	1	1	1
7. Uneven inheritance	Uneven inheritance/more genes from one parent	0.5714	0.6667	0.5	4	3
4. Gender/Age	Vague differences (Age, Gender)	0.2857	0.25	0.3333	3	4
15. Different traits from parents	Different traits/phenotype come from each parent	0.7272	0.8	0.6667	6	5
9. Genes similar	Genes/dna/traits are similar but not exactly the same	0.625	0.625	0.625	8	8
6. Same parents	Siblings have the same parents	0.875	0.7778	1	7	9
8. Same genes	Siblings share the same genes/dna/chromosomes	0.5	0.381	0.7273	11	21
3. Different genes	Siblings have different DNA/Genes/Chromosomes	0.5641	0.4583	0.7333	15	24
Off-Task						
2. Repeats question	Repeats the question without adding a response	0	0	0	0	1
Accurate ideas						
19. Difft Phenotype	Siblings have different phenotypes	0.3529	0.75	0.2308	13	4
24. 50/50 inherit	Half genes from each parent	0.6667	1	0.5	2	1
25. Punnett	Describe how a Punnett Square works	0	0	0	0	0
23. Geno/Phenotype	Genotype affects phenotype	0.375	0.3	0.5	6	10
16. Chance	Chance/Randomness of inherited genes	1	1	1	5	5

17. Dom/Recessive	Dominant vs. recessive genes	0.8889	1	0.8	5	4
18. Allele combos	Everyone has different combinations of alleles	0.72	0.6429	0.8182	11	14
20. Genes from parents	Genes/DNA/traits are inherited from parents	0.8276	0.7826	0.8788	41	46

Table 2

Average Number of Ideas per Student

Timepoint	Average Number of Ideas/Student
Pretest to Posttest	
Pretest response	2.16
Revised posttest explanation	2.67 [a]
Within-Unit Dialog	
Initial response	1.76
Revised response	2.43 [a]
Posttest Dialog	
Initial response	1.83
Revised response	2.67 [a]

Notes. [a] Significant change from previous time point.

Table 3

Frequencies of Ideas in Percentages of Students who Expressed or Added the Idea (Absolute Numbers) across the Unit

Idea	Total on Pretest	Total on Revised Posttest Answer	Change Pretest to Revised Posttest Answer	Total on Initial Within-Unit Answer	Total on Revised Within-Unit Answer	Change Within-Unit Dialog (Initial to Revised)	Total on Initial Posttest Answer	Total on Revised Posttest Answer	Change Posttest Dialog (Initial to Revised)
Vague and Off-Task Ideas									
Vague differences (e.g. age, born different times)	7.21 (44)	2.3 (14)	-4.92 (-30) [a]	3.61 (22)	3.77 (23)	0.16 (1)	2.3 (14)	2.3 (14)	0 (0)
Siblings have the same parents	19.34 (118)	17.05 (104)	-2.3 (-14)	12.62 (77)	15.41 (94)	2.79 (17)	12.62 (77)	17.05 (104)	4.43 (27) [a]
Different traits come from each parent	6.89 (42)	5.41 (33)	-1.48 (-9)	3.93 (24)	3.77 (23)	-0.16 (-1)	5.9 (36)	5.41 (33)	-0.49 (-3)
Genes / DNA similar but not same	8.2 (50)	6.89 (42)	-1.31 (-8)	7.21 (44)	8.36 (51)	1.15 (7)	8.03 (49)	6.89 (42)	-1.15 (-7)
Everyone is different	0.16 (1)	0.16 (1)	0 (0)	0.16 (1)	0.82 (5)	0.66 (4)	0.33 (2)	0.16 (1)	-0.16 (-1)
Uneven inheritance	2.95 (18)	3.11 (19)	0.16 (1)	3.28 (20)	2.13 (13)	-1.15 (-7)	1.48 (9)	3.11 (19)	1.64 (10) [a]
Siblings share same genes / DNA	26.23 (160)	27.38 (167)	1.15 (7)	19.67 (120)	28.69 (175)	9.02 (55) [a]	21.15 (129)	27.38 (167)	6.23 (38) [a]
Repeats question [off-task]	3.11 (19)	4.26 (26)	1.15 (7)	3.77 (23)	1.48 (9)	-2.3 (-14) [a]	4.26 (26)	4.26 (26)	0 (0)
Siblings have different DNA	28.03 (171)	30.66 (187)	2.62 (16)	31.97 (195)	29.67 (181)	-2.3 (-14)	29.34 (179)	30.66 (187)	1.31 (8)
Accurate Ideas									

Idea	Total on Pretest	Total on Revised Posttest Answer	Change Pretest to Revised Posttest Answer	Total on Initial Within-Unit Answer	Total on Revised Within-Unit Answer	Change Within-Unit Dialog (Initial to Revised)	Total on Initial Posttest Answer	Total on Revised Posttest Answer	Change Posttest Dialog (Initial to Revised)
Different combinations of alleles	15.74 (96)	30.33 (185)	14.59 (89) [a]	17.54 (107)	24.59 (150)	7.05 (43) [a]	20.66 (126)	30.33 (185)	9.67 (59) [a]
Genotype affects phenotype	13.61 (83)	25.9 (158)	12.3 (75) [a]	8.85 (54)	23.28 (142)	14.43 (88) [a]	10.66 (65)	25.9 (158)	15.25 (93) [a]
Randomness / Chance of inheriting genotype	9.84 (60)	17.05 (104)	7.21 (44) [a]	7.05 (43)	15.08 (92)	8.03 (49) [a]	5.74 (35)	17.05 (104)	11.31 (69) [a]
Inheritance from parents	44.1 (269)	50.82 (310)	6.72 (41) [a]	40.82 (249)	46.39 (283)	5.57 (34) [a]	40.49 (247)	50.82 (310)	10.33 (63) [a]
Dominant vs recessive	11.31 (69)	17.21 (105)	5.9 (36) [a]	4.1 (25)	14.43 (88)	10.33 (63) [a]	7.54 (46)	17.21 (105)	9.67 (59) [a]
Punnett Square	6.72 (41)	10.49 (64)	3.77 (23) [a]	2.79 (17)	9.67 (59)	6.89 (42) [a]	3.77 (23)	10.49 (64)	6.72 (41) [a]
Half genes from each parent	8.2 (50)	11.48 (70)	3.28 (20)	2.62 (16)	9.34 (57)	6.72 (41) [a]	5.25 (32)	11.48 (70)	6.23 (38) [a]
Different phenotypes	3.93 (24)	6.56 (40)	2.62 (16)	5.9 (36)	6.39 (39)	0.49 (3)	3.77 (23)	6.56 (40)	2.79 (17)
ALL vague ideas / number of students	1.02 (623)	0.97 (593)	-0.05 (-30)	0.86 (526)	0.94 (574)	0.07 (48) [a]	0.85 (521)	0.97 (593)	0.11 (73) [a]
ALL accurate ideas / number of student	1.13 (692)	1.79 (1036)	0.56 (344) [a]	0.94 (574)	1.50 (910)	0.60 (363) [a]	0.97 (597)	1.79 (1036)	0.72 (439) [a]

Notes. [a] Significant change.

Table 4

Regression of Change in KI on Changes in Ideas (Total)

Parameter	Pretest to Initial Within-Unit Dialog	Initial to Second Revision Within-Unit Dialog	Second to Third Revision Within-Unit Dialog	Third to Final Revision Within-Unit Dialog	Final Revision Within-Unit Dialog to Initial Posttest Dialog	Initial to Second Revision Posttest Dialog	Second to Third Revision Posttest Dialog	Third to Final Revision Posttest Dialog
Intercept	0.01 (0.03)	-0.04 (0.03)	0.02 (0.03)	0.15 (0.04)**	-0.23 (0.03)**	0.12 (0.03)**	0 (0.03)	0.13 (0.05)**
b	0.29 (0.02)**	0.26 (0.02)**	0.34 (0.03)**	0.27 (0.02)**	0.25 (0.02)**	0.29 (0.02)**	0.39 (0.03)**	0.26 (0.02)**
beta	0.55 (0.03)**	0.4 (0.04)**	0.44 (0.04)**	0.49 (0.04)**	0.5 (0.04)**	0.47 (0.04)**	0.47 (0.04)**	0.46 (0.04)**
R Square	0.3	0.15	0.18	0.25	0.25	0.22	0.23	0.21

Table 5

Regression of Change in KI on Changes in Ideas (Accurate)

Parameter	Pretest to Initial Within-Unit Dialog	Initial to Second Revision Within-Unit Dialog	Second to Third Revision Within-Unit Dialog	Third to Final Revision Within-Unit Dialog	Final Revision Within-Unit Dialog to Initial Posttest Dialog	Initial to Second Revision Posttest Dialog	Second to Third Revision Posttest Dialog	Third to Final Revision Posttest Dialog
Intercept	-0.04 (0.03)	-0.15 (0.03)**	0 (0.03)	0.25 (0.03)**	-0.2 (0.03)**	-0.03 (0.03)	-0.02 (0.03)	0.22 (0.04)**
b	0.41 (0.02)**	0.38 (0.03)**	0.41 (0.03)**	0.38 (0.02)**	0.35 (0.02)**	0.35 (0.03)**	0.46 (0.03)**	0.36 (0.02)**
beta	0.65 (0.03)**	0.48 (0.04)**	0.45 (0.04)**	0.58 (0.03)**	0.59 (0.03)**	0.49 (0.04)**	0.5 (0.04)**	0.55 (0.03)**
R Square	0.43	0.22	0.2	0.35	0.35	0.24	0.25	0.32

Table 6

Regression of Change in KI on Changes in Ideas (Vague)

Parameter	Pretest to Initial Within-Unit Dialog	Initial to Second Revision Within-Unit Dialog	Second to Third Revision Within-Unit Dialog	Third to Final Revision Within-Unit Dialog	Final Revision Within-Unit Dialog to Initial Posttest Dialog	Initial to Second Revision Posttest Dialog	Second to Third Revision Posttest Dialog	Third to Final Revision Posttest Dialog
Intercept	-0.18 (0.04)**	-0.15 (0.04)**	0.05 (0.04)	0.51 (0.05)**	-0.41 (0.04)**	0.03 (0.04)	-0.03 (0.04)	0.54 (0.05)**
b	-0.01 (0.03)	0.04 (0.04)	0.04 (0.06)	-0.01 (0.04)	0 (0.04)	0.08 (0.04)*	0.05 (0.06)	-0.05 (0.04)
beta	-0.01 (0.04)	0.04 (0.04)	0.03 (0.04)	-0.01 (0.04)	0 (0.04)	0.08 (0.04)*	0.04 (0.04)	-0.05 (0.04)
R Square	0	0	0	0	0	0.01	0	0

Table 7

Regression of Change in KI on Changes in Ideas (Off-task)

Parameter	Pretest to Initial Within-Unit Dialog	Initial to Second Revision Within-Unit Dialog	Second to Third Revision Within-Unit Dialog	Third to Final Revision Within-Unit Dialog	Final Revision Within-Unit Dialog to Initial Posttest Dialog	Initial to Second Revision Posttest Dialog	Second to Third Revision Posttest Dialog	Third to Final Revision Posttest Dialog
Intercept	-0.18 (0.03)**	-0.17 (0.03)**	0.5 (0.04)**	-	-0.41 (0.04)**	-0.01 (0.03)	0.51 (0.04)**	-
b	-0.07 (0.14)	-0.17 (0.19)	-0.07 (0.34)	-	-0.03 (0.16)	0.03 (0.16)	-0.07 (0.19)	-

beta	-0.02 (0.04)	-0.04 (0.04)	-0.01 (0.05)	-	-0.01 (0.04)	0.01 (0.04)	-0.02 (0.04)	-
R Square	0	0	0	-	0	0	0	-

Table 8

% Changes in Frequency of Ideas for Students who Made Different Progress in KI Scores between Different Timepoints

Idea	Pretest to Revised Posttest		Within-Unit Dialog		Posttest Dialog	
	KI 1/2 → KI 3	KI 1/2/3 → KI 4/5	KI 1/2 → KI 3	KI 1/2/3 → KI 4/5	KI 1/2 → KI 3	KI 1/2/3 → KI 4/5
Vague or Off-Task Ideas						
Repeats question	4.24	-4.35	-3.19	-4.44	3.12	-6.59 [a]
Siblings have different DNA	4.24	8.70	1.06	11.11	-3.12	10.99
Vague differences (e.g. age, born different times)	-14.41 [a]	-2.17	-4.26	0.00	-2.34	1.10
Everyone is different	-0.85	0.00	-1.06	1.11	-0.78	0.00
Siblings have the same parents	-7.63	-10.87	8.51	8.89	-0.78	6.59
Uneven inheritance	5.08	-8.70	0.00	2.22	4.69	1.10
Siblings share same genes/DNA	0.00	2.17	4.26	14.44 [a]	4.69	4.40
Genes/DNA similar but not same	-5.08	-4.35	-1.06	-1.11	-4.69	-8.79 [a]
Different traits come from each parent	4.24	-8.70	2.13	-2.22	2.34	-1.10
Accurate Ideas						
Randomness/Chance of inheriting genotype	5.08	50.00 [a]	13.83 [a]	48.89 [a]	10.16 [a]	48.35 [a]
Dominant vs recessive	10.17 [a]	34.78 [a]	10.64 [a]	42.22 [a]	7.03 [a]	38.46 [a]
Different combinations of alleles	25.42 [a]	39.13 [a]	23.40 [a]	15.56 [a]	15.62 [a]	28.57 [a]

Idea	Pretest to Revised Posttest		Within-Unit Dialog		Posttest Dialog	
	KI 1/2 → KI 3	KI 1/2/3 → KI 4/5	KI 1/2 → KI 3	KI 1/2/3 → KI 4/5	KI 1/2 → KI 3	KI 1/2/3 → KI 4/5
Different phenotypes	4.24	10.87	2.13	5.56	5.47 [a]	9.89 [a]
Inheritance from parents	51.69 [a]	15.22	43.62 [a]	27.78 [a]	50.00 [a]	24.18 [a]
Genotype affects phenotype	6.78	23.91 [a]	12.77 [a]	35.56 [a]	15.62 [a]	32.97 [a]
Half genes from each parent	3.39	30.43 [a]	5.32	33.33 [a]	4.69 [a]	40.66 [a]
Punnett Square	6.78 [a]	23.91 [a]	7.45 [a]	38.89 [a]	1.56	32.97 [a]

Note. [a] Significant change; *N* (from KI 1 or 2 (Pretest) to KI 3 (Revised Posttest Answer)) = 118, *N* (from KI 1, 2 or 3 (Prestest) to KI 4 or 5 (Revised Posttest Answer)) = 46, *N* (from KI 1 or 2 (Initial Within-Unit Answer) to KI 3 (Revised Within-Unit Answer)) = 94, *N* (from KI 1, 2 or 3 (Initial Within-Unit Answer) to KI 4 or 5 (Revised Within-Unit Answer)) = 90, *N* (from KI 1 or 2 (Initial Posttest Answer) to KI 3 (Revised Posttest Answer)) = 128, *N* (from KI 1, 2 or 3 (Initial Posttest Answer) to KI 4 or 5 (Revised Posttest Answer)) = 91.