education sciences

MDPI

Article

# A Comparison of Responsive and General Guidance to Promote Learning in an Online Science Dialog

**Libby Gerard [1,*], Marcia C. Linn [1] and Marlen Holtmann [2]**

[1] School of Education, UC Berkeley, Berkeley Way West Building (BWW), 2121 Berkeley Way, Berkeley, CA 94720-1670, USA; mclinn@berkeley.edu

[2] International Association for the Evaluation of Educational Achievement, Überseering 27, 22297 Hamburg, Germany; holtmann.marlen@gmail.com

* Correspondence: libbygerard@berkeley.edu

**Abstract:** Students benefit from dialogs about their explanations of complex scientific phenomena, and middle school science teachers cannot realistically provide all the guidance they need. We study ways to extend generative teacher–student dialogs to more students by using AI tools. We compare Responsive web-based dialogs to General web-based dialogs by evaluating the ideas students add and the quality of their revised explanations. We designed the General guidance to motivate and encourage students to revise their explanations, similar to how an experienced classroom teacher might instruct the class. We designed the Responsive guidance to emulate a student–teacher dialog, based on studies of experienced teachers guiding individual students. The analyses comparing the Responsive and the General condition are based on a randomized assignment of a total sample of 507 pre-college students. These students were taught by five different teachers in four schools. A significantly higher proportion of students added new accurate ideas in the Responsive condition compared to the General condition during the dialog. This research shows that by using NLP to identify ideas and assign guidance, students can broaden and refine their ideas. Responsive guidance, inspired by how experienced teachers guide individual students, is more valuable than General guidance.

**Keywords:** knowledge integration; technology; guidance; machine learning; AI; natural language processing; science learning; curriculum design; science teaching

check for updates

**Citation:** Gerard, L.; Linn, M.C.; Holtmann, M. A Comparison of Responsive and General Guidance to Promote Learning in an Online Science Dialog. *Educ. Sci.* **2024**, *14*, 1383. https://doi.org/10.3390/educsci14121383

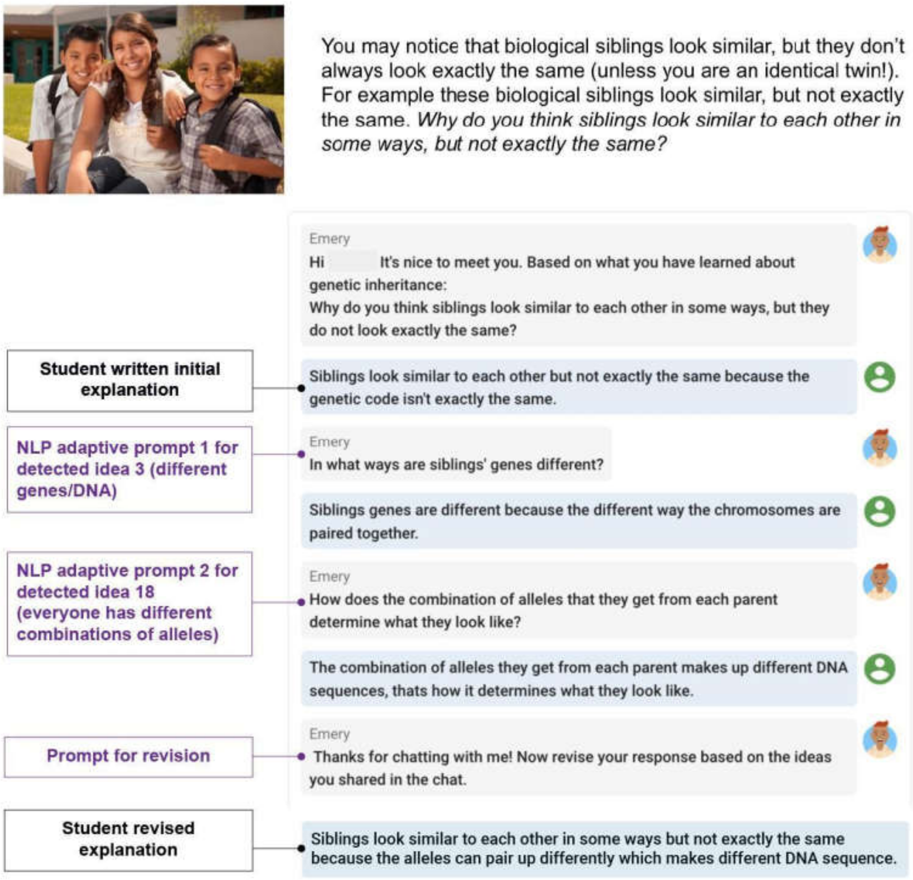
Academic Editor: Randall S. Davies

Received: 4 September 2024
Revised: 5 December 2024
Accepted: 11 December 2024
Published: 17 December 2024

## 1. Introduction

Guiding students to generate explanations during instruction and revise their reasoning by increasing coherence supports inquiry science learning. We explore ways to design guidance that scaffolds students to analyze and revise their explanations. Pedagogical research on teacher guidance shows that students revise their reasoning when they reflect on their understanding and consider how to connect their ideas with evidence [1,2]. Students benefit from personalized guidance that features a question or activity that builds on the student's idea [3]. However, highly specific guidance and guidance providing the answer may have limited or only short-term benefits [4,5]. In this study, we test a design for Responsive guidance to motivate students to revise their explanations for a Knowledge Integration question about genetic inheritance. Responsive guidance leverages NLP techniques to detect student ideas in an online science dialog and prompt the student to elaborate their idea [6]. A chatbot, referred to as a Thought Buddy, builds on the NLP-detected idea, to encourage the student to discuss the mechanisms underlying their detected idea (Figure 1).

We compare the Responsive dialog to a General dialog in which the guidance prompts the student to reflect on their explanation and add details, or distinguish which ideas they hold offer explanatory power. Both Guidance conditions align with Knowledge Integration (KI) pedagogy that emphasizes showing respect for students as science learners and encouraging them to distinguish between their ideas [7] (see Table 1 for sample detected ideas and guidance prompts). Responsive guidance emulates a student–teacher dialog, based on studies of experienced teachers guiding individual students. General guidance emulates ways an experienced teacher might instruct the whole class by encouraging students to refine their responses and does *not* align to each student's specific expressed idea. For example, the General guidance prompt in the first round of the online dialog is, "Can you tell me why you think that? Adding more details can help to clarify your thinking". Findings have implications for refining KI pedagogy and for designing NLP tools that reinforce teacher interactions with their students and improve student learning opportunities.



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?*

Emery
Hi [redacted] It's nice to meet you. Based on what you have learned about genetic inheritance:
Why do you think siblings look similar to each other in some ways, but they do not look exactly the same?

Siblings look similar to each other but not exactly the same because the genetic code isn't exactly the same.

Emery
In what ways are siblings' genes different?

Siblings genes are different because the different way the chromosomes are paired together.

Emery
How does the combination of alleles that they get from each parent determine what they look like?

The combination of alleles they get from each parent makes up different DNA sequences, thats how it determines what they look like.

Emery
Thanks for chatting with me! Now revise your response based on the ideas you shared in the chat.

Siblings look similar to each other in some ways but not exactly the same because the alleles can pair up differently which makes different DNA sequence.

Student written initial explanation
NLP adaptive prompt 1 for detected idea 3 (different genes/DNA)
NLP adaptive prompt 2 for detected idea 18 (everyone has different combinations of alleles)
Prompt for revision
Student revised explanation

**Figure 1.** Responsive dialog.

**Table 1.** Sample detectable ideas and guidance prompt (11).

| 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, but not the exact same DNA code".<br>"I think siblings look similar to each other but not the same because there are so many different ways DNA can be combined, so no one will look exactly the same as their sibling". | 1: In what ways do siblings have different genetic information?<br>2: Why do siblings look more alike than people who are not related? |

**Table 1.** *Cont.*

| ID | Detectable Idea | Sample Student Responses (Idea Detected Underlined) | Guidance Prompts (Prompt 2 Assigned If Idea Detected Twice) |
|---|---|---|---|
| 4 | Vague differences (ex: time born, gender) | "Because they can get different alleles and environmental pressures, but their dna is still similar. also, they could be different ages or gender".<br>"they look similar because they have anelle genes from each parent and not born at the same time so they all have different affects on them". | 1: One sister has dimples and the other does not. Why/How do you think this happened?<br>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 formed in the same amniotic sac".<br>"the genotypes in the same parents go to the kids which makes them similar". | 1: Why does it matter that siblings have the same parents?<br>2: Why does it matter that the parents are the same? |
| **Sample Accurate Ideas** | | | |
| 16 | Chance/Randomness | "They inherit different traits from their parents because sometimes there can be a 50% chance a child will inherit a trait from a parent".<br>"Because of the half genes (from the parents) that gets passed down are random. This means that the siblings don't necessary have to inherit the same traits". | 1: How are the genes you inherit randomly selected?<br>2: How might you predict what traits a child might have? |
| 18 | Everyone has different combinations of alleles | "there are so many different ways DNA can be combined, 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".<br>"the genes get mixed" | 1: How does the combination of alleles that they get from each parent determine what they look like?<br>2: How could you predict what combination of alleles a sibling would get from their parents? |

We investigate two research questions:

1. Does Responsive guidance (that prompts a student to analyze *their* idea) spur more students to add more accurate ideas to answer a KI item than General guidance, when embedded in an online dialog in an inquiry science unit?
2. What are the characteristics of effective guidance prompts?

## 2. Theoretical Framework: Knowledge Integration

We used the Knowledge Integration (KI) framework to inform the design of the curriculum, NLP modeling, guidance, and assessment [7,8]. The KI framework articulates how learners develop coherent and personally relevant understanding of complex science topics, based on over four decades of classroom research employing longitudinal and instructional comparison study methods. Based on this view of learning, KI articulates instructional design tenets. KI describes that learners have developed many ideas about science phenomena from their experiences outside of school and from prior classes [9,10]. Students develop more robust scientific understanding by elaborating and refining these initial ideas that are rooted in their experiences and intuitions rather than discarding them [11].

KI items elicit these ideas as a starting point. Students often do not realize that the ideas they develop from experiences outside of school and prior instruction are valuable for sense-making inside the science classroom. Eliciting a student's ideas calls for asking questions that connect to their experiences and their way of interpreting the problem, which provides the student an entry-point to articulate the evidence they are using to inform their idea [12]. Consistent with the principles of culturally relevant pedagogy, this means recognizing the scientific reasoning in a student's experientially based idea and offering scaffolds to build from this idea (e.g., refs. [13,14]). Guiding students to recognize one of their ideas as a tool for sense-making often precipitates their generation of related ideas.

A second tenet of the KI framework is to provide students with opportunities to discover new ideas, which may come from re-examining and rethinking their own ideas as well as analyzing evidence from instruction, exchanging ideas with peers, or talking to a teacher. When students re-examine evidence from instruction, guidance that prompts students to consider how that evidence connects to or challenges their initial idea has shown to have more long-term impact than guidance that tells the student to add a specific idea from the evidence [5].

Inquiry instruction often stops after eliciting students' ideas and giving guidance to discover a new idea [1]. KI points to the need for students to distinguish among their ideas to select which ones are grounded in evidence and relevant to explaining the problem. KI has identified the importance of distinguishing ideas for developing coherent understanding in studies of activities that engage students in sorting their ideas, critiquing their explanations, and analyzing peers' responses in online discussion [8]. Without opportunities to distinguish among their ideas and select promising ideas, students tend to accumulate a set of isolated ideas [15].

Another tenet of the KI framework emphasizes that students need opportunities to reflect and form links among their selected ideas to create a coherent explanation. Studies demonstrate that providing opportunities to reflect and reconsider ideas during science learning impacts outcomes [16,17]. Students benefit from guidance to revise their initial explanation that models how to connect new ideas or evidence with one's initial ideas [3].

### *Dialog Design*

We use the KI pedagogical framework to design an asset-based online dialog that positions each student as holding valuable, culturally shaped ideas from which to build more coherent scientific knowledge [18]. The structure of the embedded dialog featuring an avatar as a sense-making partner who recognizes and responds to the student's idea, rather than supplying the student with correct ideas, frames the student as holding epistemic authority. It models science as a process of reflection and revision, rather than accumulation of settled ideas [19]. For example, the NLP idea detection model detects vague ideas that are born from students' experiences in the world and asks a question to encourage the student to elaborate and refine their reasoning, positioning their idea as a productive starting point in constructing scientific knowledge. To enable a KI-based dialog, we designed NLP idea detection models that recognize each student's distinct scientific idea within their explanations for a personally relevant science problem embedded in a web-based inquiry science investigation [6].

The NLP idea detection rubric is developed based on an analysis of over 1000 student responses to a KI question (Why do biological siblings look similar but not identical?) from students who are of the same age and at the same schools or schools within the same region as the students who use the online dialog. This ensures that the training data for the idea detection model includes culturally-based ideas and linguistic variations that are closer in nature to the ideas expressed by the students who use the dialog than training data collected from a more distant sample. This increases the likelihood that the focal students' ideas can be detected by the NLP model and engaged in discussion in the online dialog.

The online dialog guidance prompts, to respond to each student's detected idea, are designed by experienced science teachers who have previously taught the unit in their classrooms. The goal is to extend their asset-based pedagogy to guide students' explanation refinement to more students, for teachers who typically have 30–35 students per class. The guidance prompts are reviewed by additional teachers and education researchers to increase the likelihood that the guidance is accessible to students, responsive to each student's ideas, and a reasonable next step to encourage the student to deepen their reasoning.

## 3. Designing Guidance

### *3.1. Teacher Guidance for Revision*

Generating and refining scientific explanations during inquiry increases science learning [12,20,21]. Scientists have repeatedly recognized the role of revising scientific explanations as crucial for incorporating experimental refinements and scientific discoveries [19,22]. Students benefit from guidance in constructing explanations and determining how to revise. Furtak et al. [23] conducted a meta-analysis of 37 experimental and quasi-experimental studies on inquiry-based science teaching from 1996 to 2006. Studies involving teacher guidance during inquiry had mean effect sizes about 0.40 larger than those with student-led conditions.

Expert teachers demonstrate how to provide effective guidance as students investigate KI problems. KI problems, such as the problem featured in this study, "Why do biological siblings look similar but not identical?" call for students to link multiple ideas [24]. Students often hold varied ideas from their experiences and from instruction, such as "they have the same parents", "they are not born at the same time", "they get more genes from one parent", or "there are different ways DNA is combined". Students gather new ideas from interacting with dynamic visualizations embedded in the unit (e.g., testing allele combinations and resulting phenotypes) and distinguish among these new ideas and their initial views. They form links among selected ideas to generate an explanation. Teacher guidance can help students to refine their initial explanations to make them more coherent [25,26].

Research illustrates effective teacher guidance designs. A multiple case study analysis of 57 lessons of 19 science teachers demonstrated that teacher guidance which prompted students to examine evidence relevant to their idea, or pressed students to further elaborate their ideas, led to sustained sense-making and greater learning compared to guidance focused on correcting students [2]. Similarly, Haverly et al. [27] found that when novice teachers "made space" for students to reflect on their understanding, such as by having another student revoice and respond to a peer′s inaccurate idea—rather than providing guidance to correct a student—students engaged in more complex and sustained sense-making during class discussions. A core tenet in designing guidance that creates space for sense-making is the teacher′s ability to recognize students' multiple ways of interpreting science phenomena as intellectually generative [18] and press for further use of evidence to refine their ideas [28].

Ruiz-Primo and Furtak [1] studied how teachers engaged students in informal assessment conversations during instruction. Assessment conversations entailed eliciting student thinking, recognizing how a student's idea contributes to understanding the topic, and then using the information about the student's idea to formulate a question or an activity that helps the student elaborate or probe their idea. They analyzed three teachers' assessment conversations and found that while student outcomes were similar across the teachers at baseline, the students of the teacher who most frequently engaged students in complete assessment conversations demonstrated significantly greater student learning gains than students of the other two teachers. These studies together point to the strengths of teacher guidance which recognizes the value of a student's initial idea and prompts the student to work with their idea, by elaborating or comparing using related evidence.

### *3.2. Automated Guidance to Emulate Teacher Guidance*

Building on the positive effects of teacher guidance for student learning, researchers have explored ways to design computer-based, automated guidance that can promote knowledge integration during inquiry learning. Automated guidance may support teachers to extend the kind of personalized guidance they want to give, to recognize a student's idea and encourage further reflection and analysis, to more students during instruction [29,30]. It can support teachers to ensure 30+ students receive an approximation of their guidance during a 50-min period. Further, it may give teachers who are unfamiliar with a science topic and teaching it for the first time assistance in interpreting a student's reasoning and a response that will push them to delve into related evidence [31]. Many teachers are teaching a topic for the first time due to changes in curriculum and standards, and the frequent reassignment of teachers to new grades or schools.

Research on automated guidance highlights dilemmas in how to design guidance to foster robust student learning [3,32]. Findings consistently indicate that overly specific guidance can be misleading by interrupting a student's line of reasoning or encouraging students to view science learning as a superficial accumulation of facts. Adaptive knowledge integration guidance, which directed students to analyze a dimension of evidence that was relevant to their initial idea, was more effective than adaptive specific guidance, which directed students to add a specific idea that was missing from their explanation [5]. Relatedly, general reflection prompts that encouraged students to "stop and think" during a science lesson were more effective than directive prompts that gave students a hint for what they should think about next in the specific instructional context [4]. Students who received the general reflection prompts developed more coherent science understanding than students who received the directive prompts over the course of the lesson. Authors argued that the general reflection prompts may have allowed and encouraged students to take control of their own reflection and hence let them build from their idea, whereas the directive prompts may not have aligned with the student's line of reasoning and hence interfered with their learning. The directive prompts notably were not adaptive, so it is less likely that they connected to the student's initial idea(s).

Studies point to the benefits of automated, personalized guidance, which encourages students to reflect on evidence relevant to their idea [33]. Adaptive guidance responsive to the students' idea was more effective than non-adaptive, randomly assigned guidance [34]. Relatedly, teacher-assigned knowledge integration guidance was more effective than teacher-assigned generic guidance (e.g., add evidence) [35]. A meta-analysis of automated adaptive guidance found that adaptive guidance which prompted reflection or elaboration of one's idea was more effective than guidance that provided corrective feedback [36]. Adaptive KI guidance coupled with an Annotator, a model of how to revise a science explanation by linking a new idea with one's initial ideas, was more effective than multiple rounds of adaptive KI guidance [3]. This shows the benefits of guidance that offers a disciplinary hint to examine related evidence and support for students to reflect on how to link their initial and new ideas together. These approaches to guidance design emulate the guidance of expert teachers, in that they demonstrate value for the students' initial idea and scaffold the student to take ownership to pursue evidence to further clarify or elaborate their initial idea.

In this study, we explore the dilemma of designing guidance that is personalized to respond to the student's idea but is not overly specific such that it precludes a student's reasoning. We compare two guidance designs within an online dialog: In the Responsive condition, the avatar uses NLP to ask the student questions that are tailored to the specific ideas the student expressed. In the General condition, the avatar's questions are pre-authored to elicit additional reasoning and are not tailored to the student's specific response. In both conditions, the guidance is grounded in the KI framework, intentionally designed to elicit the student's further elaboration of their ideas, and to encourage them to distinguish among their ideas. In the Responsive condition, the prompts use NLP to suggest elaborations to the student's specific idea, whereas in the general condition the prompts do not use NLP and press generally for reflection and elaboration.

## 4. Methods

### *4.1. Participants*

The analyses comparing the Responsive and the General condition are based on a randomized assignment of a total sample of 507 pre-college students (267 students in the Responsive condition and 240 students in the General condition). Each student was randomly assigned by the web-based learning environment to one of the two conditions in the unit executed using the branching tool in the learning environment. These students were taught by five different teachers in four schools. School demographic data covers a range of student demographics(students who identify as belonging to non-white racial groups: 89–54%; students eligible for free or reduced-price lunch: 69–10%; students labeled as English Language Learners: 13–3%). To obtain a richer database for analyzing the prompts, we added 681 pre-college students from seven schools that taught the genetics unit with the siblings dialog. The study was approved by University[X]'s Institutional Review Board for research with human subjects.

### *4.2. Siblings Adaptive Dialog in a Genetic Inheritance Curriculum*

To design the dialog, we developed NLP idea detection models for a constructed response item named 'Siblings: Why do siblings look similar but not exactly the same?' The question is embedded in a Web-based Inquiry Science Environment (WISE)) unit on genetic inheritance. The prompt elicits links among student ideas about the inheritance of genes from each biological parent and genetic variation. It encourages students to draw on their personal experience or observations of siblings and parents' appearances and ideas from the genetics unit to form a response.

Students encounter the Siblings question in the WISE unit. They choose an avatar from a group of images that look like students or robots. The avatar then greets the student by name and asks them to share their ideas about the Siblings prompt. The student writes their response. The student then engages in two rounds of conversation with the avatar, in which the avatar asks a question, the student responds, the avatar asks another question, and the student responds. Then the student is prompted to use what they learned from the conversation to revise their initial explanation to Siblings.

In the Responsive condition, the avatar uses NLP to ask the student questions that are tailored to the specific ideas the student expressed (Table 1). In the General condition, the avatar's questions are pre-authored to elicit additional reasoning and are not tailored to the student's specific response. The first general prompt asks, "Can you tell me why you think that? Adding more details can help to clarify your thinking". The second asks, "Of all your ideas, which one best explains your thinking about why siblings look similar but not exactly the same?".

### *4.3. NLP Idea Detection*

#### 4.3.1. Training Data

A dataset of 1485 student explanations in response to the Siblings KI prompt was collected from prior research with schools whose school demographics are similar to the schools that used the NLP models in this study.

#### 4.3.2. Idea Detection Rubric

Two teachers (one former biology teacher and one current biology teacher) who are a part of our research group analyzed approximately 300 of the student responses in the training dataset to create an idea detection rubric that was comprehensive of the student ideas about genetic inheritance within the dataset (Table 1). An idea was typically made up of a phrase within a sentence (e.g., underlined segments in Table 1). The teachers set the boundaries for an idea by identifying expressions within a student's explanation that they would 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. Each student response could include zero (e.g., "IDK"), one, or more ideas.

The initial idea detection rubric enumerated the distinct ideas emergent from the students' answers in the training dataset grouped in terms of accurate ideas and vague 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 raters reapplied the updated rubric to a new 10% data sample to ensure the refined version captured all expressed student ideas. The 10% sample provides a sufficient quantity of student responses for two human raters to compute inter-rater reliability. The final version of the Siblings idea detection rubric identified 25 distinct ideas that were found in students' responses.

#### 4.3.3. Annotation

The two teachers annotated the ideas in all of the student responses in the training dataset, using INCEpTION [37]. The teachers trained on the data until they achieved inter-rater reliability for idea-detection. Once reaching inter-rater reliability for idea detection, each teacher annotated 50% of the student responses in the dataset by applying the idea detection rubric. They assigned a tag for each distinct idea within each response.

#### 4.3.4. NLP Modeling for Idea Detection

A computer scientist used the teacher-annotated training data to build the NLP idea detection model [6,11,38]. 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 [39] were used for idea detection. Transformer model training involved starting from a large pre-trained model and fine-tuning the model weights on training examples specific to idea detection [40,41]. See [11] 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 [38], the idea detection model and adaptive dialog were embedded into the online Genetic Inheritance unit.

Researchers integrate the NLP service into the web-based unit. Each teacher creates a run of the unit for their class, which creates a copy of the main unit. Each of their students registers in the learning environment and uses an access code to add the unit. When a student submits a response to an idea-detection dialog in the unit, the system sends the student response to the NLP service. The service runs the response through the NLP model and sends a result object back that includes the ideas detected. The learning environment parses the result and uses the information to assign guidance. Guidance is pre-authored for each detectable idea (Table 1). The whole process of detecting ideas and assigning guidance takes 1–2 s. This process occurs for two rounds of dialog guidance. Then, the student is prompted to revise their initial explanation based on what they learned in the dialog. The learning environment logs all responses.

### *4.4. Data*

#### 4.4.1. Within-Unit Adaptive Dialog

The dialog was embedded within the Genetic Inheritance unit after students completed a lesson about genetic inheritance and Punnett squares in which they explored 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.

#### 4.4.2. Post-Test Adaptive Dialog

Students completed the Siblings item with the dialog again on the unit post-test after completing the unit. This was approximately twelve days after they completed the within-unit dialog. Between the within-unit dialog and the post-test dialog, students completed additional activities about genetics.

In the within-unit and post-test adaptive dialogs, the initial prompt for the Siblings item was the same in both conditions. In the Responsive condition, students responded to different adaptive prompts based on the ideas detected in their answers. In the General condition, students responded to the same prompts within each dialog. The last prompt in the within-unit and post-test dialog was the same, asking students to revise their initial response

#### 4.4.3. Logged Analytics

WISE logged 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. The ideas for each response were exported into anonymized csv data files for analysis.

#### 4.4.4. Student Feedback

Immediately after completing the dialog within the unit and at post-test, students were asked a multiple-choice question: How did the Thought Buddy interpret your answers? (a) The Thought Buddy understood my ideas; (b) The Thought Buddy did not understand my ideas. Explain your choice.

### *4.5. Data Analysis*

For all analyses, we used the software R (v4.2.1 [42]).

#### 4.5.1. RQ1: Impacts of Responsive vs. General Guidance on Student Learning

***Student Adding of Ideas***. We compared student data at three different time points between the Responsive and the General condition:

1. First round of dialog: From initial explanation to the first response in the dialog.
2. Second round of dialog: From the first to the second response in the dialog.
3. Revision of explanation: From their initial explanation at the start of the dialog, to their final explanation after the dialog.

We were interested in what proportion of students added at least one new accurate idea at each time point, what proportion of students added at least one vague idea, and how many students repeated one of the accurate ideas they had mentioned in their previous answer. To test if there is a significant difference between the proportions of the two independent conditions, we used the Two-Proportion Z-Test.

We further explored the differences of students' responses between the two conditions by comparing the proportion of how often the different detected ideas were added within each condition. This proportion highlights which ideas were added more or less often during the dialog within each condition. We again conducted the Two-Proportion Z-Test to test if the proportion of students that added each idea within each condition significantly differed.

***Student Reports on Guidance***. We wanted to know if there was a significant difference between conditions in the frequency of students who reported that the Thought Buddy understood their ideas. We coded student reports as 1 (understood my ideas) or 0 (did not). To test if there is a significant difference between the proportion of students in the two independent conditions, we used the Two-Proportion Z-Test.

#### 4.5.2. RQ2: Characteristics of Effective Guidance Prompts

For RQ2, we combined the data from the 507 students who participated in this comparison study with data from an additional 681 pre-college students who also used the genetics unit with the Siblings dialog and NLP-based responsive guidance in a previous study. This provided a larger dataset to analyze the effectiveness of the prompts. We evaluated the prompts based on whether they supported students to add a new, accurate idea across the time points. The proportion of students who added an accurate idea after the previous prompt indicates the success rate of prompts, i.e., how many of these students added a new accurate idea after this prompt. Additionally, since the prompt was selected based on students' responses before, we added the absolute frequency of students who added a new accurate idea after the prompt. This allowed us to identify the prompts that were used frequently and had a high (or low) success rate. Furthermore, we were interested in the ideas that were triggered by these most successful prompts. This enabled us to observe whether these prompts tended to encourage students to develop accurate mechanistic or vague ideas.

## 5. Results

### *5.1. RQ1: Does Responsive Guidance (That Prompts a Student to Analyze Their Idea) Spur More Students to Add Accurate Ideas to Answer a KI Item Than General Guidance, When Embedded in an Online Inquiry Dialog in an Inquiry Science Unit?*

#### 5.1.1. Frequency of Ideas Added

The Responsive dialog encouraged students to add new accurate ideas more frequently than the General dialog. During the first and second round of the within-unit dialog, 72% of the students in the Responsive condition added at least one new accurate idea in comparison to their initial explanation, while 59% of the students in the General condition did so (Table 2). The proportion of students in the post-test dialog was lower (47% in the Responsive condition and 37% in the General condition), and showed a similar pattern.

**Table 2.** Comparison of the frequency and proportion of students who added at least one new accurate/vague idea.

| Guidance | Count | Proportion | Timepoint | Idea |
|---|---|---|---|---|
| adaptive | 222 | 83% | Within-unit dialog | Accurate |
| generic | 174 | 72% | Within-unit dialog | Accurate |
| adaptive | 157 | 59% | Post-test dialog | Accurate |
| generic | 113 | 47% | Post-test dialog | Accurate |
| adaptive | 192 | 72% | First and Second Round of Within-Unit Dialog | Accurate |
| generic | 142 | 59% | First and Second Round of Within-Unit Dialog | Accurate |
| adaptive | 125 | 47% | First and Second Round of Post-test Dialog | Accurate |
| generic | 89 | 37% | First and Second Round of Post-test Dialog | Accurate |
| adaptive | 149 | 56% | Within-unit dialog | Vague |
| generic | 161 | 67% | Within-unit dialog | Vague |
| adaptive | 99 | 37% | Post-test dialog | Vague |
| generic | 91 | 38% | Post-test dialog | Vague |

The Responsive prompts had the largest effect compared to the General prompts during the second round of the dialog (Figure 2). In the within-unit dialog and in the post-test dialog, from the first to the second round of the dialog, a significantly higher proportion of students in the Responsive condition compared to the General condition added at least one new accurate idea.

There were no other significant differences between the two conditions. Although students in the Responsive condition were more likely to add new accurate ideas during the dialog, they did not integrate more new ideas when they revised their explanation after the dialog than the students in the General condition. This suggests that the Responsive prompts were more effective in helping students to add new accurate ideas to their repertoire. More work is needed to assist students in integrating their new ideas.

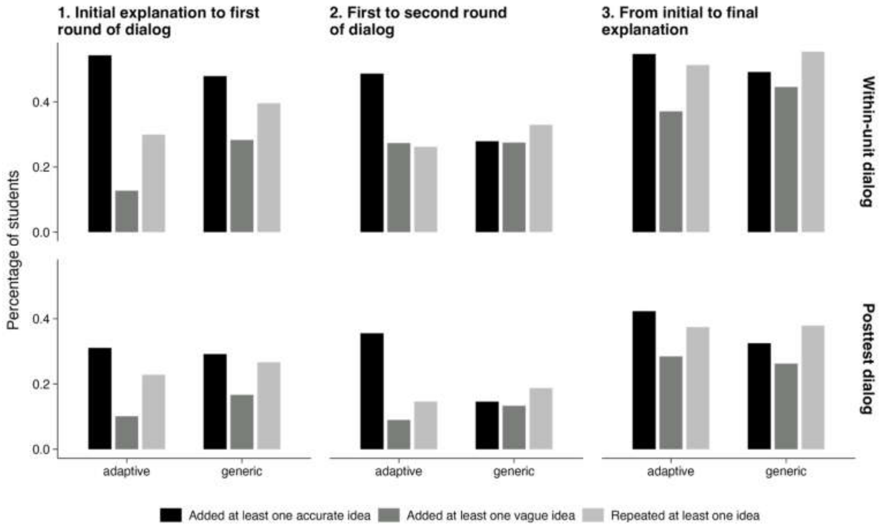


**Figure 2.** Proportion of students who added a new accurate idea, a new vague idea or repeated at least one idea of their previous response at each time point.

### 5.1.2. Types of Ideas Added

We investigated the addition of accurate, mechanistic ideas in general and also identified the actual ideas added.

***Initial explanation to the first round of dialog***. Students in the Responsive condition were more likely to add accurate ideas that helped explain the mechanisms of genetic inheritance during both the within-unit and post-test dialogs, compared to students in the General condition. A significantly higher proportion of students in the Responsive condition than in the General condition added the idea of *dominant* vs. *recessive alleles* from their initial explanation to the first round of both dialogs (Figure 3). Students in the Responsive condition also added the idea that *half of one's genes come from each parent* more often than students in the General condition from the initial explanation to the first round of the within-unit dialog.

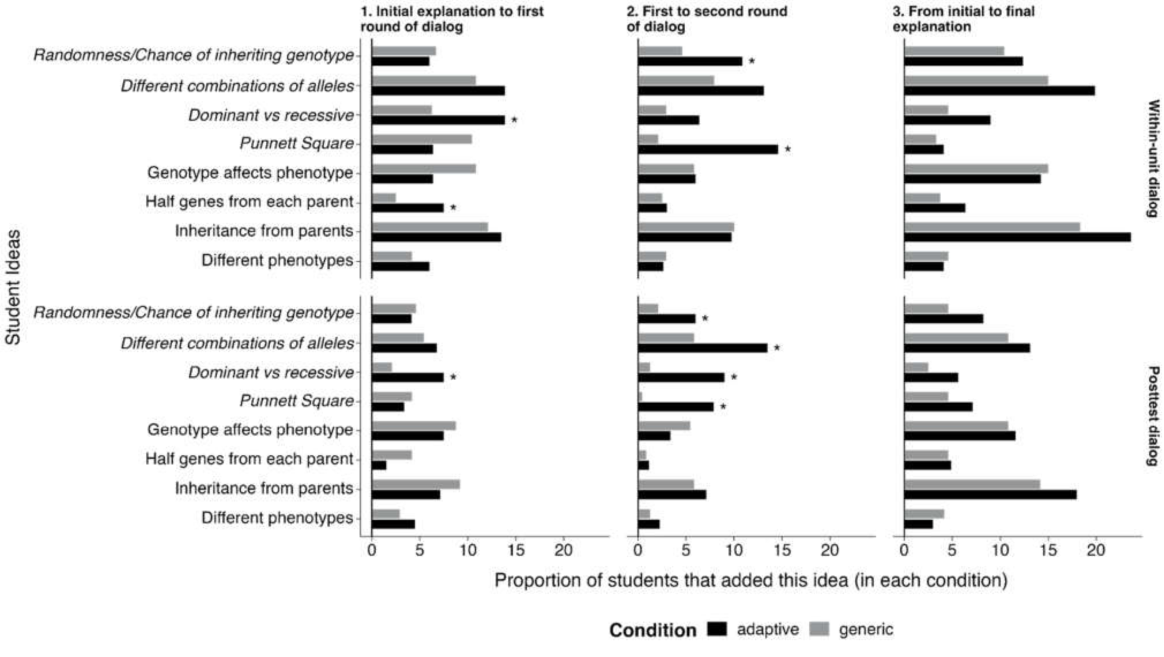


**Figure 3.** The proportion of students in each condition who added this idea. Notes. *Mechanistic ideas are in italics. * indicates p < 0.05.*

***First to second round of the dialog***. The second round of the dialog encouraged more students in the Responsive than in the General condition to express mechanistic ideas about genetic inheritance (Figure 2; Appendix A, Tables A1 and A2). A higher proportion of students in the Responsive condition added the idea about the *randomness of inheriting a genotype* and the *use of a Punnett Square* to explain genetic inheritance than in the General condition in both the within-unit and post-test dialogs. In addition, a significantly higher proportion of students in the Responsive condition added two more accurate ideas that hold explanatory power for genetic inheritance: *different combinations of alleles* and *dominant* vs. *recessive alleles*.

***Initial to final explanation***. When students revised their explanations after the dialog, they added similar types of ideas whether they were in the Responsive or General condition. The differences in the types of ideas students added when compared to the ideas in their initial explanation were not significant in the within-unit or post-test dialogs.

### 5.1.3. Student Opinions About Responsive Versus General Guidance

The curriculum asked students immediately after they engaged in the online dialog if they thought the avatar understood their ideas. A total of 69% of the students in the Responsive condition reported that the avatar "understood their ideas", compared to 58% of students in the General condition. This difference in proportions is significant with a $p$-value of $p = 0.01$.

This pattern was repeated in the post-test dialog. A total of 52% of the students in the Responsive condition reported that the avatar understood their ideas, compared to 36% of the students in the General condition. This difference in proportions is significant with a $p$-value $< 0.001$.

## *5.2. RQ2: What Are the Characteristics of Effective Guidance Prompts?*

To identify effective guidance prompts, we analyzed how students responded to each prompt in each round of the dialog. We computed the proportion of students who added at least one new accurate idea when they received the prompt, in comparison to the ideas in their previous responses.

In the first round of the dialog, 47% of the students receiving the General prompt added an idea (Table 3). Eight Responsive guidance prompts elicited more new accurate ideas than the General prompt. In the second round of the dialog, only 27% of the students added a new idea after the General guidance prompt. Fifteen prompts in the Responsive condition were more effective (Table 4). In sum, the most successful prompts in both rounds were in the Responsive condition, echoing the benefits reported in results for Research Question 1. This indicates that guidance connected to a student's idea was more likely to assist the student in stating a new accurate idea than the General guidance, particularly in the second round of the dialog.

**Table 3.** Successful prompts in the first round of the dialog.

| Prompt | Frequency of the Prompt | Proportion of Students Who Added an Accurate Idea After the Prompt | Number of Students Who Added an Accurate Idea After the Prompt |
|---|---|---|---|
| How do children get different allele combinations? | 52 | 81% | 42 |
| **Tell me more about how siblings get some different genes**. | 65 | 70% | 45 |
| How does the combination of alleles that they get from each parent determine what they look like? | 170 | 66% | 112 |
| **In what ways do siblings have different genetic information?** | 67 | 55% | 37 |

**Table 3.** *Cont.*

| Prompt | Frequency of the Prompt | Proportion of Students Who Added an Accurate Idea After the Prompt | Number of Students Who Added an Accurate Idea After the Prompt |
|---|---|---|---|
| How are the genes you inherit randomly selected? | 51 | 52% | 27 |
| How do the genes children get from their parents determine which traits they'll have? | 119 | 50% | 59 |
| How do the genes each sibling inherits from their parents determine their traits? | 135 | 49% | 67 |
| How can the combination of alleles make one sibling look different from the other? | 68 | 49% | 33 |
| Can you tell me why you think that? Adding more details can help to clarify your thinking. | 396 | 47% | 187 |

Prompts with a frequency lower than 15 not shown. The grayed-out prompt is the prompt of the General condition. **Bold** guidance is for vague ideas; plain text is guidance for accurate ideas.

**Table 4.** Successful prompts in the second round of the dialogs.

| Prompt | Frequency of the Prompt | Proportion of Students Who Added an Accurate Idea After the Prompt | Number of Students Who Added an Accurate Idea After the Prompt |
|---|---|---|---|
| How does the combination of alleles that they get from each parent determine what they look like? | 19 | 88% | 16 |
| How could you predict what combination of alleles a sibling would get from their parents? | 121 | 80% | 97 |
| How do children get different allele combinations? | 17 | 78% | 13 |
| How might you predict what traits a child might have? | 39 | 74% | 29 |
| How does the combination of genes from each parent affect the traits each sibling inherits? | 141 | 66% | 93 |
| How can the combination of alleles make one sibling look different from the other? | 35 | 62% | 21 |
| In what ways are siblings' genes different? | 15 | 58% | 8 |
| Why don't both siblings end up with the same combination of dominant and recessive alleles? | 101 | 57% | 57 |
| How does each sibling's allele combination determine how similar they look to one another? | 80 | 52% | 41 |
| **Why does it matter that the parents are the same?** | 70 | 49% | 33 |
| *How can two parents produce the same or different traits for each sibling? Check out LINK and then come back and tell me what you find out!* | 308 | 48% | 146 |
| **How does genetic material make siblings look more alike than people who are not related?** | 21 | 47% | 10 |
| Since children inherit genes from their parents, why don't siblings always have the same traits? | 163 | 39% | 65 |
| How do the genes each sibling inherits from their parents determine their traits? | 48 | 36% | 19 |
| Why don't siblings always have the same traits as their parents? | 101 | 35% | 35 |
| Of all your ideas, which one best explains your thinking about why siblings look similar but not exactly the same? | 371 | 27% | 103 |

Prompts with a frequency lower than 15 not shown. The grayed-out prompt is the prompt of the General condition. **Bold** guidance is for vague ideas; plain text is guidance for accurate ideas; *italicized* prompts are guidance for undetectable ideas or 'I don't know' responses.

Among the top five most successful prompts across the two rounds of the dialog, three of the prompts were elaborating, supporting students to pursue different lines of reasoning and two were deepening, supporting students to analyze an idea in greater depth (Table 5). Across all students, eight different accurate ideas were added after the elaborating guidance prompts, with no more than 30% of the students adding one of the ideas. In contrast, the two successful deepening prompts narrowed the majority of the students' thinking to one idea, the Punnett Square, with about 45% of students adding this idea in response to the guidance. The two deepening guidance prompts used the terminology 'how would you predict' which may have led students to narrow in on the idea of a Punnett Square.

**Table 5.** Which ideas were elicited by the most successful prompts?

| **Idea** | **How Do Children Get Different Allele Combinations? (Round 1)** | **Tell Me More About How Siblings Get Some Different Genes. (Round 1)** | **How Does the Combination of Alleles That They Get from Each Parent Determine What They Look Like? (Round 1)** | **How Could You Predict What Combination of Alleles a Sibling Would Get from Their Parents? (Round 2)** | **How Might You Predict What Traits a Child Might Have? (Round 2)** | **How Does the Combination of Genes from Each Parent Affect the Traits Each Sibling Inherits? (Round 2)** |
|---|---|---|---|---|---|---|
| Randomness/Chance of inheriting genotype | 9<br>(9%) | 15<br>(14%) | 17<br>(7%) | 23<br>(13%) | 6<br>(11%) | 28<br>(14%) |
| Dominant vs. recessive | 13<br>(14%) | 14<br>(13%) | 55<br>(23%) | 16<br>(9%) | 4<br>(7%) | 28<br>(14%) |
| Different combinations of alleles | 23<br>(24%) | 10<br>(9%) | 26<br>(11%) | 21<br>(12%) | 6<br>(11%) | 47<br>(23%) |
| Different phenotypes | 6<br>(6%) | 6<br>(6%) | 21<br>(9%) | 2<br>(1%) | 3<br>(5%) | 12<br>(6%) |
| Inheritance from parents | 27<br>(28%) | 35<br>(33%) | 42<br>(17%) | 18<br>(10%) | 7<br>(13%) | 37<br>(18%) |
| Genotype affects phenotype | 6<br>(6%) | 6<br>(6%) | 52<br>(21%) | 11<br>(6%) | 2<br>(4%) | 16<br>(8%) |
| Half genes from each parent | 6<br>(6%) | 8<br>(7%) | 9<br>(4%) | 4<br>(2%) | 1<br>(2%) | 13<br>(6%) |
| Punnett Square | 5<br>(5%) | 13<br>(12%) | 20<br>(8%) | 77<br>(45%) | 26<br>(48%) | 25<br>(12%) |

The Responsive prompt for students who stated 'I don't know' or a similar expression (e.g., IDK; random letters) encourages students to re-examine evidence within the unit. Thus, when the NLP model could not detect an idea that it was trained on, or when it detected that a student said, "I don't know", a prompt was assigned that said, "How can two parents produce the same or different traits for each sibling? Check out '**link**' and then come back and tell me what you find out!". This guidance was used frequently (308 times in the second round of the dialog) and helped almost half of the students to express an accurate idea in the next round of the dialog.

Responsive guidance was more beneficial for students who started with an initially accurate idea than for students who started with a vague idea. Among the prompts that resulted in over 25% of respondents adding a new accurate idea, two-thirds of these guidance prompts were assigned to students who expressed an accurate idea in their initial explanation. However, when a student expressed this vague idea in their explanation, '*Siblings' genes are similar but not identical*', the prompt 'Tell me more about how siblings get some different genes' was highly successful in guiding students to articulate a new accurate idea, resulting in over 70% of students adding a new accurate idea. The prompt that most frequently elicited the new idea '*genes are inherited from parents*' (Table 4) was demonstrated in prior research to be a productive starting point for learning of genetic inheritance [11]).

## 6. Discussion

Generating and revising scientific explanations during learning contributes to developing robust and coherent knowledge (e.g., ref. [12]). We explored how web-based dialogs leveraging NLP can support teachers to engage each student in a class of 30–35 students in a dialog to elaborate and refine their science explanations of genetic inheritance. NLP-based Responsive guidance was more effective than General guidance in supporting students to generate new, accurate scientific ideas when engaged in dialog. Responsive guidance was designed to emulate effective teacher guidance designs that support students to recognize their idea as a valuable starting point and pursue a new idea by analyzing their idea and related evidence to elaborate or refine their idea. General guidance was designed to resemble the guidance teachers give to the whole class to encourage thoughtful responses. These findings add detailed insights to a growing body of research demonstrating that NLP can be leveraged to provide guidance that effectively supports teachers in fostering students' knowledge integration in science [3,5].

The findings add to our understanding of how Responsive guidance benefits students' learning. This study highlights that Responsive guidance supported more students than in the General guidance condition to generate new accurate scientific ideas, compared to the ideas in their initial explanation. Responsive guidance was personalized to prompt the student to analyze the evidence and reasoning underlying their initial idea. It did not provide corrective feedback to the student's idea or point the student to a specific idea to add. This suggests that students have more ideas in their repertoire to explain a scientific dilemma than they initially recognize as relevant, and/or that students generate new ideas by analyzing their initial ideas.

The results show that students benefit from guidance which helps them recognize the value of their idea as a starting point, and a question that helps them analyze the details and evidence behind their idea to identify or generate a new, related idea. This echoes the benefits of effective teacher guidance which presses students to elaborate their ideas [2]. As one student in the Responsive condition said, "When I put in my ideas they [Thought Buddy] understood and gave me more to go off of". Another remarked, "he [Thought Buddy] made me think about my idea".

Three of the most successful guidance prompts resulted in students elaborating their response by adding multiple, different accurate ideas to extend their reasoning. Two of the most successful prompts guided students to deepen their initial idea. The guidance created a productive space for students to pursue their next step in a line of scientific reasoning, adapting to the nature of the idea held by the student. The guidance expanded the possible pathways in classroom science learning, consistent with the goal of respecting each student's reasoning while encouraging them to consider all of the ideas available. This extends Davis' [4] findings that reflection prompts were more helpful for students than (non-adaptive) contextualized hints which, by giving students a specific direction to pursue, may inadvertently have interfered with students' direction of reasoning. Leveraging NLP allows for designing guidance that combines a prompt for reflection with idea-level adaptivity, to spur reconsideration of the evidence and ideas one started with.

Reassuringly, the students in the Responsive condition compared to the General condition were more likely to report that the avatar understood their ideas. Some reported that 'the avatar looked like them', and 'sounded like a normal person'. Students' reflections suggest that students do discern when guidance is designed to connect to their science ideas and presence in the classroom. This aligns with prior research showing that students are more motivated to learn when they perceive instruction to be adapted for them [43].

Further design-based research is needed to understand how to best guide students who begin with vague ideas to take advantage of an accurate idea and elaborate their response. We also need further insight into how to leverage the NLP idea detection to support students in linking the ideas they raise in a dialog with their initial ideas. Students raised more accurate new ideas in the dialog in the Responsive condition, yet the rates of adding new ideas into their revised explanations were similar between the Responsive and General guidance conditions. This suggests that while students in the Responsive condition generated new accurate ideas, they did not integrate them to reformulate a more robust and coherent understanding.

## 7. Practical Implications for Science Teaching

This research has three major practical implications for science teachers. First, adaptive dialogs can expand teachers' awareness of the science ideas their students have. Students are often reluctant to share their science ideas in class discussions. This may be due to concern that their idea is incorrect or irrelevant. The adaptive dialog elicited additional ideas students held, which students had not expressed in their initial science explanation. Successful guidance prompts elicited different, accurate ideas. By guiding students to analyze the evidence underlying their initial ideas, the adaptive dialog, similar to an expert teacher, sparked students to recognize and express more of the science ideas they held. Seeing their students' ideas in the dialogs may expand teachers' knowledge of the diversity and quantity of the ideas their students hold about the science topic. Expanded knowledge of student ideas can support teachers to affirm their students as science learners and to provide responsive instruction.

Second, the adaptive dialogs can assist teachers in providing personalized guidance for each student to revise their written science explanations within classes of 30–35 students. While students are engaging with the computer-based adaptive dialog, the teacher is freed to circulate the classroom to assist students who are stuck, or to extend students' conversations with their avatar to discussions with their peers. Or, teachers may lead a class discussion to highlight the different ideas emerging across student dialogs in the class to communicate the value of distinguishing among multiple perspectives in building science knowledge. Designing adaptive dialogs to support teachers in providing guidance can ensure that each student has their ideas recognized and the opportunity to extend their explanation.

Last, the web-based unit with adaptive dialog featured in this study is open-source, meaning it and others are freely available and customizable using the learning environment's authoring and customization environment. Teachers may try this unit or other units with their students, and customize the pre-authored adaptive guidance for each detectable idea in the NLP-based dialogs to better align with their own tested classroom guidance strategies.

## 8. Conclusions

In conclusion, this research shows that by using NLP to identify ideas and assign guidance emulating that used by experienced teachers, students can broaden and refine their ideas. Responsive guidance, inspired by how experienced teachers guide individual students, is more valuable than General guidance inspired by how experienced teachers instruct their whole class.

**Author Contributions:** Conceptualization, L.G., M.C.L. and M.H. Methodology: L.G., M.C.L. and M.H.; Software, M.H.; Formal analysis, M.H.; Investigation, L.G. and M.C.L.; Writing—original draft, L.G.; Writing—review and editing, L.G., M.C.L. and M.H.; Project administration, M.C.L. and L.G.; Funding acquisition, M.C.L. and L.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** National Science Foundation, Grant Number 2101669.

**Institutional Review Board Statement:** Committee for Protection of Human Subjects, Institutional Review Board, UC Berkeley Technologies to Inform Practice in Science (TIPS), 2021-06-14389.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The datasets presented in this article are not readily available because of the policies of the Institutional Review Board.

**Acknowledgments:** The authors gratefully acknowledge the partnership of the science teachers and the researchers who contributed to the NLP model development, guidance design, classroom teaching, and the interpretation of the results.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Appendix A

**Table A1.** Percentage of students who added at least one accurate or one vague idea or repeated at least one idea.

| Timepoint | Condition | Percentage of Students Who Added at Least One Accurate Idea | Percentage of Students Who Added at Least One Vague Idea | Percentage of Students Who Repeated at Least One Idea |
|---|---|---|---|---|
| 1. Initial explanation to first round of dialog | adaptive | 0.31 | 0.10 | 0.23 |
| 1. Initial explanation to first round of dialog | generic | 0.29 | 0.17 | 0.27 |
| 1. Initial explanation to first round of dialog | adaptive | 0.54 | 0.13 | 0.30 |
| 1. Initial explanation to first round of dialog | generic | 0.48 | 0.28 | 0.40 |
| 2. First to second round of dialog | adaptive | 0.36 | 0.09 | 0.15 |
| 2. First to second round of dialog | generic | 0.15 | 0.13 | 0.19 |
| 2. First to second round of dialog | adaptive | 0.49 | 0.27 | 0.26 |
| 2. First to second round of dialog | generic | 0.28 | 0.28 | 0.33 |
| 3. From initial to final explanation | adaptive | 0.42 | 0.28 | 0.37 |
| 3. From initial to final explanation | generic | 0.32 | 0.26 | 0.38 |
| 3. From initial to final explanation | adaptive | 0.55 | 0.37 | 0.51 |
| 3. From initial to final explanation | generic | 0.49 | 0.45 | 0.55 |

**Table A2.** The frequency and percentage of adding the idea at each time point and condition.

| Guidance | Pattern | Idea | Frequency | Percentage |
|---|---|---|---|---|
| adaptive | 1. Initial explanation to first round of dialog | Randomness/Chance of inheriting genotype | 16 | 8.16 |
| adaptive | 1. Initial explanation to first round of dialog | Dominant vs. recessive | 37 ** | 18.88 ** |
| adaptive | 1. Initial explanation to first round of dialog | Different combinations of alleles | 37 | 18.88 |
| adaptive | 1. Initial explanation to first round of dialog | Different phenotypes | 16 | 8.16 |
| adaptive | 1. Initial explanation to first round of dialog | Inheritance from parents | 36 | 18.37 |
| adaptive | 1. Initial explanation to first round of dialog | Genotype affects phenotype | 17 | 8.67 ** |
| adaptive | 1. Initial explanation to first round of dialog | Half genes from each parent | 20 * | 10.20 ** |
| adaptive | 1. Initial explanation to first round of dialog | Punnett Square | 17 | 8.67 ** |
| adaptive | 2. First to second round of dialog | Randomness/Chance of inheriting genotype | 29 * | 16.38 |
| adaptive | 2. First to second round of dialog | Dominant vs. recessive | 17 | 9.60 |
| adaptive | 2. First to second round of dialog | Different combinations of alleles | 35 | 19.77 |

**Table A2.** *Cont.*

| Guidance | Pattern | Idea | Frequency | Percentage |
|---|---|---|---|---|
| adaptive | 2. First to second round of dialog | Different phenotypes | 7 | 3.95 |
| adaptive | 2. First to second round of dialog | Inheritance from parents | 26 | 14.69 ** |
| adaptive | 2. First to second round of dialog | Genotype affects phenotype | 16 | 9.04 * |
| adaptive | 2. First to second round of dialog | Half genes from each parent | 8 | 4.52 |
| adaptive | 2. First to second round of dialog | Punnett Square | 39 ** | 22.03 ** |
| adaptive | 2. First to second round of dialog | Randomness/Chance of inheriting genotype | 33 | 13.20 |
| adaptive | 3. From initial to final explanation | Dominant vs. recessive | 24 | 9.60 |
| adaptive | 3. From initial to final explanation | Different combinations of alleles | 53 | 21.20 |
| adaptive | 3. From initial to final explanation | Different phenotypes | 11 | 4.40 |
| adaptive | 3. From initial to final explanation | Inheritance from parents | 63 | 25.20 |
| adaptive | 3. From initial to final explanation | Genotype affects phenotype | 38 | 15.20 |
| adaptive | 3. From initial to final explanation | Half genes from each parent | 17 | 6.80 |
| adaptive | 3. From initial to final explanation | Punnett Square | 11 | 4.40 |
| adaptive | 1. Initial explanation to first round of dialog | Randomness/Chance of inheriting genotype | 11 | 9.73 |
| adaptive | 1. Initial explanation to first round of dialog | Dominant vs. recessive | 20 ** | 17.70 ** |
| adaptive | 1. Initial explanation to first round of dialog | Different combinations of alleles | 18 | 15.93 |
| adaptive | 1. Initial explanation to first round of dialog | Different phenotypes | 12 | 10.62 |
| adaptive | 1. Initial explanation to first round of dialog | Inheritance from parents | 19 | 16.81 |
| adaptive | 1. Initial explanation to first round of dialog | Genotype affects phenotype | 20 | 17.70 |
| adaptive | 1. Initial explanation to first round of dialog | Half genes from each parent | 4 | 3.54 ** |
| adaptive | 1. Initial explanation to first round of dialog | Punnett Square | 9 | 7.96 |
| adaptive | 2. First to second round of dialog | Randomness/Chance of inheriting genotype | 16 * | 11.94 |
| adaptive | 2. First to second round of dialog | Dominant vs. recessive | 24 ** | 17.91 ** |
| adaptive | 2. First to second round of dialog | Different combinations of alleles | 36 ** | 26.87 |
| adaptive | 2. First to second round of dialog | Different phenotypes | 6 | 4.48 ** |
| adaptive | 2. First to second round of dialog | Inheritance from parents | 19 | 14.18 ** |
| adaptive | 2. First to second round of dialog | Genotype affects phenotype | 9 | 6.72 ** |
| adaptive | 2. First to second round of dialog | Half genes from each parent | 3 | 2.24 |
| adaptive | 2. First to second round of dialog | Punnett Square | 21 ** | 15.67 ** |
| adaptive | 3. From initial to final explanation | Randomness/Chance of inheriting genotype | 22 | 11.52 |
| adaptive | 3. From initial to final explanation | Dominant vs. recessive | 15 | 7.85 |
| adaptive | 3. From initial to final explanation | Different combinations of alleles | 35 | 18.32 |
| adaptive | 3. From initial to final explanation | Different phenotypes | 8 | 4.19 |
| adaptive | 3. From initial to final explanation | Inheritance from parents | 48 | 25.13 |

**Table A2.** *Cont.*

| Guidance | Pattern | Idea | Frequency | Percentage |
|---|---|---|---|---|
| adaptive | 3. From initial to final explanation | Genotype affects phenotype | 31 | 16.23 |
| adaptive | 3. From initial to final explanation | Half genes from each parent | 13 | 6.81 |
| adaptive | 3. From initial to final explanation | Punnett Square | 19 | 9.95 |
| generic | 1. Initial explanation to first round of dialog | Randomness/Chance of inheriting genotype | 16 | 10.46 |
| generic | 1. Initial explanation to first round of dialog | Dominant vs. recessive | 15 ** | 9.80 ** |
| generic | 1. Initial explanation to first round of dialog | Different combinations of alleles | 26 | 16.99 |
| generic | 1. Initial explanation to first round of dialog | Different phenotypes | 10 | 6.54 |
| generic | 1. Initial explanation to first round of dialog | Inheritance from parents | 29 | 18.95 |
| generic | 1. Initial explanation to first round of dialog | Genotype affects phenotype | 26 | 16.99 ** |
| generic | 1. Initial explanation to first round of dialog | Half genes from each parent | 6 * | 3.92 ** |
| generic | 1. Initial explanation to first round of dialog | Punnett Square | 25 | 16.34 ** |
| generic | 2. First to second round of dialog | Randomness/Chance of inheriting genotype | 11 * | 11.83 |
| generic | 2. First to second round of dialog | Dominant vs. recessive | 7 | 7.53 |
| generic | 2. First to second round of dialog | Different combinations of alleles | 19 | 20.43 |
| generic | 2. First to second round of dialog | Different phenotypes | 7 | 7.53 |
| generic | 2. First to second round of dialog | Inheritance from parents | 24 | 25.81 ** |
| generic | 2. First to second round of dialog | Genotype affects phenotype | 14 | 15.05 * |
| generic | 2. First to second round of dialog | Half genes from each parent | 6 | 6.45 |
| generic | 2. First to second round of dialog | Punnett Square | 5 ** | 5.38 ** |
| generic | 3. From initial to final explanation | Randomness/Chance of inheriting genotype | 25 | 13.89 |
| generic | 3. From initial to final explanation | Dominant vs. recessive | 11 | 6.11 |
| generic | 3. From initial to final explanation | Different combinations of alleles | 36 | 20.00 |
| generic | 3. From initial to final explanation | Different phenotypes | 11 | 6.11 |
| generic | 3. From initial to final explanation | Inheritance from parents | 44 | 24.44 |
| generic | 3. From initial to final explanation | Genotype affects phenotype | 36 | 20.00 |
| generic | 3. From initial to final explanation | Half genes from each parent | 9 | 5.00 |
| generic | 3. From initial to final explanation | Punnett Square | 8 | 4.44 |
| generic | 1. Initial explanation to first round of dialog | Randomness/Chance of inheriting genotype | 11 | 11.11 |
| generic | 1. Initial explanation to first round of dialog | Dominant vs. recessive | 5 ** | 5.05 ** |
| generic | 1. Initial explanation to first round of dialog | Different combinations of alleles | 13 | 13.13 |
| generic | 1. Initial explanation to first round of dialog | Different phenotypes | 7 | 7.07 |
| generic | 1. Initial explanation to first round of dialog | Inheritance from parents | 22 | 22.22 |
| generic | 1. Initial explanation to first round of dialog | Genotype affects phenotype | 21 | 21.21 |
| generic | 1. Initial explanation to first round of dialog | Half genes from each parent | 10 | 10.10 ** |

**Table A2.** *Cont.*

| Guidance | Pattern | Idea | Frequency | Percentage |
|---|---|---|---|---|
| generic | 1. Initial explanation to first round of dialog | Punnett Square | 10 | 10.10 |
| generic | 2. First to second round of dialog | Randomness/Chance of inheriting genotype | 5 * | 9.09 |
| generic | 2. First to second round of dialog | Dominant vs. recessive | 3 ** | 5.45 ** |
| generic | 2. First to second round of dialog | Different combinations of alleles | 14 ** | 25.45 |
| generic | 2. First to second round of dialog | Different phenotypes | 3 | 5.45 |
| generic | 2. First to second round of dialog | Inheritance from parents | 14 | 25.45 ** |
| generic | 2. First to second round of dialog | Genotype affects phenotype | 13 | 23.64 ** |
| generic | 2. First to second round of dialog | Punnett Square | 1 ** | 1.82 ** |
| generic | 2. First to second round of dialog | Half genes from each parent | 2 | 3.64 |
| generic | 3. From initial to final explanation | Randomness/Chance of inheriting genotype | 11 | 8.15 |
| generic | 3. From initial to final explanation | Dominant vs. recessive | 6 | 4.44 |
| generic | 3. From initial to final explanation | Different combinations of alleles | 26 | 19.26 |
| generic | 3. From initial to final explanation | Different phenotypes | 10 | 7.41 |
| generic | 3. From initial to final explanation | Inheritance from parents | 34 | 25.19 |
| generic | 3. From initial to final explanation | Genotype affects phenotype | 26 | 19.26 |
| generic | 3. From initial to final explanation | Half genes from each parent | 11 | 8.15 |
| generic | 3. From initial to final explanation | Punnett Square | 11 | 8.15 |

Notes. ** indicates $p < 0.01$; * indicates $p < 0.05$.

## References

1. Ruiz-Primo, M.A.; Furtak, E.M. Exploring Teachers' Informal Formative Assessment Practices and Students' Understanding in the Context of Scientific Inquiry. *J. Res. Sci. Teach.* **2007**, *44*, 57–84. [CrossRef]
2. Kang, H.; Windschitl, M.; Stroupe, D.; Thompson, J. Designing, Launching, and Implementing High Quality Learning Opportunities for Students That Advance Scientific Thinking. *J. Res. Sci. Teach.* **2016**, *53*, 1316–1340. [CrossRef]
3. Gerard, L.; Linn, M.C. Computer-Based Guidance to Support Students' Revision of Their Science Explanations. *Comput. Educ.* **2022**, *176*, 104351. [CrossRef]
4. Davis, E.A. Prompting Middle School Science Students for Productive Reflection: Generic and Directed Prompts. *J. Learn. Sci.* **2003**, *12*, 91–142. [CrossRef]
5. Vitale, J.M.; McBride, E.; Linn, M.C. Distinguishing Complex Ideas about Climate Change: Knowledge Integration vs. Specific Guidance. *Int. J. Sci. Educ.* **2016**, *38*, 1548–1569. [CrossRef]
6. Riordan, B.; Bichler, S.; Bradford, A.; King Chen, J.; Wiley, K.; Gerard, L.; Linn, M.C. An Empirical Investigation of Neural Methods for Content Scoring of Science Explanations. In Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, Online, 10 July 2020; Association for Computational Linguistics; pp. 135–144.
7. Linn, M.C.; Eylon, B.-S. *Science Learning and Instruction: Taking Advantage of Technology to Promote Knowledge Integration*; Routledge: New York, NY, USA, 2011.
8. Linn, M.C.; Donnelly-Hermosillo, D.; Gerard, L. Synergies Between Learning Technologies and Learning Sciences: Promoting Equitable Secondary School Teaching. In *Handbook of Research on Science Education*; Lederman, N.G., Zeidler, D.L., Lederman, J.S., Eds.; Routledge: New York, NY, USA, 2023; pp. 447–498. ISBN 978-0-367-85575-8.
9. diSessa, A.A. *Changing Minds: Computers, Learning and Literacy*; MIT Press: Cambridge, MA, USA, 2000.
10. Scardamalia, M.; Bereiter, C. Knowledge Building: Theory, Pedagogy, and Technology. In *Cambridge Handbook of the Learning Sciences*; Sawyer, K., Ed.; Cambridge University Press: New York, NY, USA, 2006; pp. 97–118.
11. Gerard, L.; Holtmann, M.; Riordan, B.; Linn, M.C. Impact of an Adaptive Dialog That Uses Natural Language Processing to Detect Students' Ideas and Guide Knowledge Integration. *J. Educ. Psychol.* **2024**. [CrossRef]
12. Chi, M.T.H.; de Leeuw, N.; Chiu, M.; LaVancher, C. Eliciting Self-Explanations Improves Understanding. *Cogn. Sci.* **1994**, *18*, 439–477.

13. Gutiérrez, K.D. 2011 AERA Presidential Address: Designing Resilient Ecologies: Social Design Experiments and a New Social Imagination. *Educ. Res.* **2016**, *45*, 187–196. [CrossRef]
14. Morales-Doyle, D. *Transformative Science Teaching: A Catalyst for Justice and Sustainability*; Harvard University Press: Cambridge, MA, USA, 2024; ISBN 978-1-68253-874-6.
15. Clark, D.B. Longitudinal Conceptual Change in Students' Understanding of Thermal Equilibrium: An Examination of the Process of Conceptual Restructuring. *Cogn. Instr.* **2006**, *24*, 467–563. [CrossRef]
16. Chi, M.T.H.; Kang, S.; Yaghmourian, D.L. Why Students Learn More From Dialogue- Than Monologue-Videos: Analyses of Peer Interactions. *J. Learn. Sci.* **2017**, *26*, 10–50. [CrossRef]
17. Haridza, R.; Ding, L. Enhancing Students' Reasoning Skills by Asking Students to Provide Evidence-Based Explanations in Science Classrooms: Findings from TIMSS 2019. *Int. J. Sci. Educ.* **2024**, 1–26. [CrossRef]
18. Rosebery, A.S.; Warren, B.; Tucker-Raymond, E. Developing Interpretive Power in Science Teaching. *J. Res. Sci. Teach.* **2016**, *53*, 1571–1600. [CrossRef]
19. Schwartz, R.; Lederman, N. What Scientists Say: Scientists' Views of Nature of Science and Relation to Science Context. *Int. J. Sci. Educ.* **2008**, *30*, 727–771. [CrossRef]
20. Berland, L.K.; Schwarz, C.V.; Krist, C.; Kenyon, L.; Lo, A.S.; Reiser, B.J. Epistemologies in Practice: Making Scientific Practices Meaningful for Students. *J. Res. Sci. Teach.* **2016**, *53*, 1082–1112. [CrossRef]
21. Engle, R.A.; Conant, F.R. Guiding Principles for Fostering Productive Disciplinary Engagement: Explaining an Emergent Argument in a Community of Learners Classroom. *Cogn. Instr.* **2002**, *20*, 399–483. [CrossRef]
22. Sandoval, W.A.; Reiser, B.J. Explanation-Driven Inquiry: Integrating Conceptual and Epistemic Scaffolds for Scientific Inquiry. *Sci. Educ.* **2004**, *88*, 345–372. [CrossRef]
23. Furtak, E.M.; Seidel, T.; Iverson, H.; Briggs, D.C. Experimental and Quasi-Experimental Studies of Inquiry-Based Science Teaching: A Meta-Analysis. *Rev. Educ. Res.* **2012**, *82*, 300–329. [CrossRef]
24. Williams, M.; Debarger, A.H.; Montgomery, B.L.; Zhou, X.; Tate, E. Exploring Middle School Students' Conceptions of the Relationship between Genetic Inheritance and Cell Division. *Sci. Educ.* **2012**, *96*, 78–103. [CrossRef]
25. Puntambekar, S.; Stylianou, A.; Goldstein, J. Comparing Classroom Enactments of an Inquiry Curriculum: Lessons Learned From Two Teachers. *J. Learn. Sci.* **2007**, *16*, 81–130. [CrossRef]
26. Manz, E. Resistance and the Development of Scientific Practice: Designing the Mangle Into Science Instruction. *Cogn. Instr.* **2015**, *33*, 89–124. [CrossRef]
27. Haverly, C.; Calabrese Barton, A.; Schwarz, C.V.; Braaten, M. "Making Space": How Novice Teachers Create Opportunities for Equitable Sense-Making in Elementary Science. *J. Teach. Educ.* **2020**, *71*, 63–79. [CrossRef]
28. Davis, E.A.; Palincsar, A.S. Engagement in High-Leverage Science Teaching Practices among Novice Elementary Teachers. *Sci. Educ.* **2023**, *107*, 291–332. [CrossRef]
29. Zhai, X.; Yin, Y.; Pellegrino, J.W.; Haudek, K.C.; Shi, L. Applying Machine Learning in Science Assessment: A Systematic Review. *Stud. Sci. Educ.* **2020**, *56*, 111–151. [CrossRef]
30. Zhai, X.; Haudek, K.C.; Shi, L.; Nehm, R.H.; Urban-Lurain, M. From Substitution to Redefinition: A Framework of Machine Learning-Based Science Assessment. *J. Res. Sci. Teach.* **2020**, *57*, 1430–1459. [CrossRef]
31. Davis, E.A.; Petish, D.; Smithey, J. Challenges New Science Teachers Face. *Rev. Educ. Res.* **2006**, *76*, 607–651. [CrossRef]
32. Koedinger, K.R.; Aleven, V. Exploring the Assistance Dilemma in Experiments with Cognitive Tutors. *Educ. Psychol. Rev.* **2007**, *19*, 239–264. [CrossRef]
33. Lee, H.-S.; Pallant, A.; Pryputniewicz, S.; Lord, T.; Mulholland, M.; Liu, O.L. Automated Text Scoring and Real-Time Adjustable Feedback: Supporting Revision of Scientific Arguments Involving Uncertainty. *Sci. Educ.* **2019**, *103*, 590–622. [CrossRef]
34. Walker, E.; Rummel, N.; Koedinger, K.R. Adaptive Intelligent Support to Improve Peer Tutoring in Algebra. *Int. J. Artif. Intell. Educ.* **2014**, *24*, 33–61. [CrossRef]
35. Gerard, L.F.; Ryoo, K.; McElhaney, K.W.; Liu, O.L.; Rafferty, A.N.; Linn, M.C. Automated Guidance for Student Inquiry. *J. Educ. Psychol.* **2016**, *108*, 60–81. [CrossRef]
36. Gerard, L.F.; Matuk, C.F.; McElhaney, K.W.; Linn, M.C. Automated, Adaptive Guidance for K-12 Education. *Educ. Res. Rev.* **2015**, *15*, 41–58. [CrossRef]
37. Klie, J.-C.; Bugert, M.; Boullosa, B.; de Castilho, R.E.; Gurevych, I. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Santa Fe, NW, USA, 23 June 2018.
38. Schulz, C.; Meyer, C.M.; Gurevych, I. Challenges in the Automatic Analysis of Students' Diagnostic Reasoning. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 6974–6981. [CrossRef]
39. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019.
40. Burstein, J.; Doran, C.; Solorio, T. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; Volume 1, pp. 4171–4186.

41. Alhindi, T.; Ghosh, D. "Sharks Are Not the Threat Humans Are": Argument Component Segmentation in School Student Essays. In Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, Online, 20 April 2021; Burstein, J., Horbach, A., Kochmar, E., Laarmann-Quante, R., Leacock, C., Madnani, N., Pilán, I., Yannakoudakis, H., Zesch, T., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 210–222.
42. *R Core Team R: A Language and Environment for Statistical Computing 2022*; R Core Team: Vienna, Austria, 2022.
43. Shute, V.J. Focus on Formative Feedback. *Rev. Educ. Res.* **2008**, *78*, 153–189. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.