Title

Designing and Customizing AI Adaptive Dialogs in Middle School Science Classrooms

ISBN

9798189270086

Author

Li, Weiying

Publication Date

2026-05-01

Designing and Customizing AI Adaptive Dialogs in Middle School Science Classrooms

By

Weiying Li

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Education

and the Designated Emphasis in

New Media

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Marcia C. Linn, Chair

Professor Michelle Honda Wilkerson

Professor Zachary A. Pardos

Spring 2026

Designing and Customizing AI Adaptive Dialogs in Middle School Science Classrooms

By

Weiying Li

ABSTRACT

Designing and Customizing AI Adaptive Dialogs in Middle School Science Classrooms

by

Weiying Li

Doctor of Philosophy in Education

Designated Emphasis in New Media

University of California, Berkeley

Professor Marcia C. Linn, Chair

This dissertation investigates how AI adaptive dialogs can be designed, iteratively refined, and customized in partnership with teachers to support middle school students' integration of complex science and social justice ideas in culturally and linguistically diverse classrooms. A central argument is that teacher expertise within a Research-Practice Partnership (RPP) operates at two distinct levels: through collective co-design, teachers shape AI dialogs that work broadly across diverse classrooms; through local customization, individual teachers amplify what those dialogs can accomplish for specific students in specific community contexts. These two levels form an iterative refinement in which local observations feed back into collective improvement, generating design knowledge that neither level could produce alone..

The Knowledge Integration (KI) framework and Justice-Centered Science Pedagogy (JCSP) provided a merged theoretical foundation for dialog design, instruction, and analysis. KI guided rubric development, adaptive guidance design, and student learning assessment through a five-level scoring rubric that rewards linked reasoning. JCSP extended KI's commitment to honoring diverse student ideas to include community observations and structural critique of systemic inequity. Together, the frameworks shaped how the RPP built the dialogs, supported teacher customization, and measured what students learned.

The dissertation draws on five years of design-based research across six socioeconomically and racially diverse schools in the western United States, involving more than 5,000 student responses analyzed through cumulative link mixed models, generalized estimating equations, and McNemar tests, alongside qualitative analysis of teacher interviews and classroom observations.

Chapter 2 establishes that iterative teacher-informed rubric expansion improved both NLP model accuracy and student KI gains across four design cycles of the photosynthesis dialog. Chapter 3 shows that two teachers' distinct pedagogical goals, making ecosystem connections visible and making science accessible for emergent bilingual learners, shaped dialog refinement and classroom instruction in complementary ways, both producing significant learning gains. Chapter 4 extends this work to food justice, showing that RPP co-design can build effective AI dialogs for

social justice science topics, and that teachers' pedagogical orientations toward food justice shaped which ideas became available in students' reasoning.

Together the three chapters show that AI dialogs become most effective when teacher expertise is embedded in the tools themselves through sustained co-design, and when teachers design instruction that treats the dialog as a starting point rather than a finished product. This dissertation offers the field actionable design guidelines and empirical evidence for building AI dialogs that honor the full range of student thinking, including community-based, culturally grounded, and justice-oriented ideas, and that are worthy of the students and teachers they are designed to serve.

Table of Contents

Acknowledgements

This dissertation grew alongside me, shaped by the generosity, wisdom, and care of so many people I am deeply grateful for.

To Marcia, thank you for your unwavering support over these six years. Our weekly meetings to workshop ideas, make sense of data, and work through findings and writing have shaped me as a researcher in ways I could not have imagined. Your mentorship has shown me what it means to truly invest in a scholar, and I am fortunate to have learned from your example.

To Michelle, thank you for your nurturing presence throughout this journey. Your writing group gave me a community of care and taught me to engage with others' work with generosity and thoughtfulness.

To Zach, thank you for introducing me to the technical foundations of this work, for teaching alongside me, and for opening your lab meetings as a space of learning and growth.

My research is deeply rooted in the TELS community, which has shaped both my scholarship and my growth as a researcher. To Libby, thank you for your generous mentorship and for walking me through so many stuck moments with patience and clarity. To Jon and Hiroko, thank you for supporting my exploration of instructional technology and for helping me build the technical skills that made this work possible.

To the teachers in this study, this dissertation would not exist without your creativity, trust, and deep care for your students. Visiting your classrooms reminded me every single time why I chose education research. I am honored to have built this work alongside you.

To my friends who walked alongside me through this PhD, especially Kelly, Meg, Aukeem, and so many others, thank you for helping me navigate life in a country far from home. You brought me laughter, gave me space to vent, and always had my back. I could not have made this journey without your friendship.

To my parents, thank you for believing in me in the way only you know how. You showed your love not only in words but in every meal you cooked when I came home, every walk we took together, and every gathering with family when you sensed I needed to feel held. Your love has carried me across every ocean and every doubt.

To Tianze, my partner, thank you for your steadfast presence through all of it. You showed up for me again and again, and you always knew I would get here, even on the days I forgot.

Chapter 1: Introduction

This dissertation is guided by the overarching research question: how can we design and customize AI adaptive dialogs with Research-Practice Partnership (RPP; Coburn & Penuel, 2016) to support middle school students' integration of complex science and social justice ideas in diverse classrooms? I examine how AI dialogs embedded in Web-based Inquiry Science Environment (WISE) units can be designed with the RPP to honor the full range of student thinking, including ideas shaped by community experience, cultural knowledge, and observations of social inequity, alongside ideas that align with scientific standards. A central argument of this dissertation is that teacher expertise within RPP operates at two distinct levels, both of which matter: through collective RPP co-design, teachers shape AI dialogs and units that work broadly across diverse classrooms; through local customization, teachers amplify what those dialogs can accomplish for specific students in specific community contexts.

My dissertation explores these questions within the context of a long-standing RPP (Linn, Clark, & Slotta, 2003; Gerard, Bradford, & Linn, 2022), which brings together middle school science teachers, educational researchers, and computer scientists to develop and refine AI-enhanced curriculum for diverse classrooms. Over four years, the RPP iteratively designed two AI adaptive dialogs embedded in a WISE school garden unit focused on photosynthesis, ecosystem, and food access. These dialogs use natural language processing (NLP) to detect students' ideas in real time and deliver personalized adaptive guidance that builds on what students already know. More details about the RPP and WISE platform are provided in the General Methods section of this chapter.

1 Research Questions

The first study (Chapter 2), "Iterative design of Photosynthesis AI Dialog with Research-Practice Partnership", is a design study examining four iterations of the photosynthesis AI dialog designed within the RPP. This chapter operates primarily at the collective RPP design level, documenting how teachers, learning scientists, and computer scientists iteratively improved both the NLP model and student learning in benchmark assessments without instruction across more than 3,400 students. Chapter 2 is guided by two research questions: What changes were made to the dialog guidance design and idea rubric across the four iterations to better support student knowledge integration, informed by teacher expertise? and What characteristics of guidance strategies in AI adaptive dialogs best support students to incorporate ideas and improve KI scores?

The second study (Chapter 3), "Impact of Science Teachers' Pedagogical Goals in Customizing an AI Adaptive Dialog to Promote Knowledge Integration in a Research-Practice Partnership," operates primarily at the classroom level. It examines how two focal teachers, Amy and Ben, used their distinct pedagogical goals to customize both AI dialog guidance and classroom practices during the photosynthesis, ecosystems, and school garden unit, and how those goal-driven customizations produced different learning outcomes. Chapter 3 is guided by two research questions: How do teachers' emerging pedagogical goals shape AI dialog customizations? and How do teachers' pedagogical goals guide their enactment of their customized dialogs and classroom practices in the photosynthesis and ecosystems unit?

The third study (Chapter 4), "Designing AI Dialogs for Food Justice" operates at both levels simultaneously. Chapter 4 is not a separate study appended to the dissertation; it is a continuation of the same design trajectory. The photosynthesis, ecosystems, and food justice unit evolved as an integrated learning environment anchored in the school garden: students first learned photosynthesis to understand plant growth, then applied that knowledge to design their school garden. After they learned the energy and matter flow within the school garden ecosystem, they investigated fresh food access in nearby neighborhoods to understand the inequitable impacts of food apartheid and to design actionable projects to make changes. Because food justice is a social justice science issue where student ideas are shaped by lived experience of structural inequity, it also poses new design demands that extend our understanding of how teacher-informed design can produce AI tools that honor community knowledge. The chapter addresses these two research questions: How does the RPP build a shared understanding of food justice reasoning and develop the food justice dialog across two iterative design cycles? How do teachers with different pedagogical orientations toward food justice enact the dialog in their classrooms?

2 Motivation

My background in psychology and prior work designing educational technology gave me frameworks for thinking about how people learn and how tools should serve them, but it was arriving in United States inquiry-based classrooms that fundamentally shifted how I understood whose knowledge counts in science education. I grew up in China, where my science education centered on mastering established content through high-stakes examinations. When I came to the United States for doctoral study, I was genuinely surprised by what I observed in classrooms: teachers asking students what they noticed and wondered about, building lessons around students' own ideas, and treating prior knowledge as something to build on rather than overcome. Students were using science to make sense of their own lives. I found myself positioned as a learner in those classrooms, listening carefully and staying curious about what I did not yet understand, and that posture became the foundation of how I approach this research. It also shaped how I approached the RPP: not as a researcher delivering solutions to teachers, but as a collaborator learning from their expertise, and designing structures that enabled all partners to learn from each other across the three years of this study.

That posture brought me into close collaboration with our partner teachers, and what I observed in those collaborations shaped the questions driving this dissertation. Working alongside Amy, one of our long-standing RPP partner teachers, I saw how deeply her knowledge of her students and their communities shaped what was possible in her science classroom. When we redesigned the photosynthesis unit to anchor it in the school garden on her campus, a place her students knew, tended, and cared about, students engaged with science in a way I had not seen before. They were not learning about ecosystems in the abstract; they were reasoning about plants they had grown, soil they had touched, and food they wanted to eat. Amy's knowledge of that garden, those students, and that neighborhood made that possible. What excited me was the prospect of designing AI adaptive dialogs with teachers like Amy to scale that kind of responsive, context-grounded teaching, so that every student could have their ideas recognized and built upon.

This dissertation asks how that kind of teacher-informed AI dialog design works, and what it makes possible. The three empirical chapters examine this question across science and social justice topics and three years of iterative co-design: how teacher expertise improves an AI dialog through RPP collaboration, how individual teachers' pedagogical goals shape what a dialog can accomplish in their specific classrooms, and whether the same approach can produce AI tools that honor students' community-based reasoning about social justice science issues. Together they build toward an answer to the question I kept returning to in Amy's classroom: what becomes possible when the people who know students best are the ones shaping the tools designed to serve them.

3 Theoretical Foundations

3.1 Knowledge Integration and Justice-Centered Science Pedagogy

The Knowledge Integration (KI) framework (Linn & Eylon, 2011) serves as the primary theoretical and analytical foundation for this dissertation. KI builds on extensive research documenting that students arrive in science classrooms with diverse ideas shaped by everyday observation, prior instruction, and lived experience (diSessa, 1993; Inhelder & Piaget, 1958). Rather than treating these ideas as misconceptions to be replaced, KI views them as productive resources: learning happens when students sort out their own ideas alongside new ones they encounter in science class, strengthening their reasoning and building more coherent understanding over time (Songer & Linn, 1991). This is guided by four instructional KI processes: eliciting student explanations, encouraging students to discover additional ideas, guiding students to use evidence to distinguish among their ideas, and asking students to reformulate links among their ideas promotes robust understanding of science concepts (Linn et al., 2023).

KI research has also surfaced four design principles that translate these commitments into instructional decisions (Kali, 2006). Making science accessible to each learner means connecting scientific concepts to the experiences and contexts that are meaningful to students, rather than relying on generic examples (Linn & Hsi, 2000). Making student thinking visible means designing activities that require students to articulate and link their ideas with evidence, so that reasoning becomes available for reflection and refinement (Ritchhart & Perkins, 2008). Helping students learn from others means structuring opportunities for students to encounter and engage with diverse perspectives, including peers' ideas (Matuk & Linn, 2023). Promoting autonomy means guiding students' reasoning rather than simply providing answers, supporting them to develop their own sense-making strategies (Davis, 2003).

KI operates at three levels in this dissertation (see Table 1 in Chapter 3 for full details). First, KI design principles provided the framework for designing the instructional materials and for understanding how teachers' pedagogical goals shaped their customization decisions. Second, KI processes provided analytic categories for coding teacher instruction during classroom practices: we identified whether teachers' instructional moves encouraged students to elicit, discover, distinguish, or connect their ideas. Third, the five-level KI scoring rubric (Ryoo & Linn, 2015) assessed student learning outcomes by measuring integrated understanding, rewarding students for creating valid links among ideas rather than penalizing misconceptions.

Justice-Centered Science Pedagogy (JCSP; Morales-Doyle, 2017) provides the second theoretical framework, anchoring the food justice work in Chapter 4 and informing the equity orientation of the dissertation as a whole. JCSP articulates a vision of science education that simultaneously honors students' ideas and community knowledge, maintains high expectations for disciplinary learning, and fosters students' critical consciousness about the relationship between science and social inequity. It is informed by culturally relevant pedagogy (Ladson-Billings, 1995) and critical pedagogy (Freire, 1970), and it positions science education as a process of enabling learners to use scientific knowledge to identify the factors contributing to inequities and generate strategies for change.

KI and JCSP share a foundational commitment rooted in sociocultural theories of learning: both begin from the premise that students arrive in classrooms with valuable prior knowledge. KI recognizes that learners hold varied ideas shaped by observation and lived experience (Matuk & Linn, 2023), and that supporting robust understanding means eliciting those ideas, providing opportunities to discover new ones, using evidence to distinguish among them, and guiding students to form connections (Linn & Eylon, 2011). A justice-centered perspective extends what KI means by making science accessible: relevance is not just about connecting science to students' everyday experiences, but about examining differential impacts and using science to take action in one's own community (Morales-Doyle, 2017). At the same time, KI's four instructional processes provide a concrete pedagogical roadmap for enacting JCSP commitments in the classroom and in AI dialog design. Eliciting honors students' community-based knowledge as a starting point. Discovering creates space to encounter structural explanations alongside personal observations. Distinguishing supports students to evaluate competing accounts of why food access is unequal. Connecting supports students to form integrated explanations that link scientific evidence to justice claims. Together, the frameworks call for instruction and AI tools that elicit and value the full range of student ideas, provide opportunities to discover community-grounded evidence, help students use science to understand structural inequities, and support students in explaining and acting toward a more just world (Bradford, Gerard, et al., 2023).

3.2 Research-Practice Partnerships and Teacher-informed Design

Research-Practice Partnership (RPP) brings together researchers and practitioners in a long-term, mutually beneficial collaboration to address shared problems in educational practice (Coburn & Penuel, 2016). The RPP within which this dissertation is situated spans more than two decades, beginning with the Computer as Laboratory Partner project, which marked the first uses of personal computers in science classrooms (Linn & Hsi, 2000). With the advent of the Internet, the RPP designed the WISE platform and refined it through iterative design studies in which teachers piloted web-based inquiry units, surfaced patterns in students' diverse ideas, and jointly negotiated revisions to curriculum and guidance across multiple years (Linn, Clark, & Slotta, 2003). Central to the partnership has always been teacher customization: the iterative adaptation of instructional materials to align with teachers' own pedagogical priorities and with the specific needs, ideas, and experiences of their students and communities (Gerard, Spitulnik, & Linn, 2010). Research has shown that teachers who participate in professional learning designed around logged student data produce stronger knowledge integration gains in their students than those who receive shorter, less contextually grounded support (Gerard et al., 2011).

As the partnership's work has extended to AI-supported adaptive dialogs, teacher-informed design has taken on new significance. Each partner in the RPP contributes distinct domain expertise: researchers lead theoretical and analytical work, teachers provide pedagogical grounding and local classroom knowledge, and computer scientists drive technical development. This structure addresses a limitation that runs through much of the AI in education literature. Systems such as ASSISTments (Heffernan & Heffernan, 2014) and OATutor (Pardos et al., 2023) have demonstrated that AI-generated help can support student learning at scale, and recent work suggests that AI-generated feedback can produce learning gains comparable to human tutor-authored support on well-defined tasks (Pardos & Bhandari, 2024). Automated scoring approaches using transformer-based models and rubric representations have also advanced the capacity to assess student responses reliably (Condor et al., 2021, 2022). However, these systems are typically designed without sustained input from the teachers who implement them or the students whose ideas they must recognize. Rubrics are fixed early and rarely revised in response to authentic classroom data. Recent work on collaborative prompt engineering, such as PromptHive (Reza et al., 2025), points in a more promising direction by bringing subject matter experts into the AI content creation process. This dissertation extends that direction by centering teachers as co-designers across the full arc of AI dialog development: from initial rubric construction and model training through iterative guidance revision and classroom implementation.

Effective teacher-informed design depends on supporting teachers to notice their students' ideas by attending not only to the scientific content of what students write but to how they express knowledge through their cultural experiences and community observations (Luna, 2018). Research on technology-mediated teacher noticing suggests that teachers' pedagogical goals shape what they attend to in student work, and that tools should be designed not only to support student reasoning but to systematically cue teachers to notice and respond to the disciplinary substance of students' thinking (Walkoe, Wilkerson, & Elby, 2017; Wilkerson et al., 2016). When teacher noticing is deliberately integrated into technology design processes, it becomes a resource for building tools that are more responsive to the diversity of student thinking.

4 General Methods, Participants, and Technologies

4.1 Methodological Approach

The studies in this dissertation employ design-based research (DBR; Sandoval & Bell, 2004) as the overarching methodological approach. DBR combines rigorous theory with systematic design and engineering to understand and address problems of practice, making it well suited to the dissertation's dual goals of improving AI dialog design and generating knowledge about how teacher-informed design works. The studies use mixed methods. Quantitative analyses include cumulative link mixed models (CLMMs) to examine ordinal KI score changes across dialog timepoints, generalized estimating equations (GEE) to model idea frequency patterns while accounting for repeated measures and student-level nesting, and Poisson GEE models for guidance revision analyses. Qualitative analyses include thematic coding of pre-implementation meeting transcripts, classroom observation fieldnotes, and post-implementation interviews, as well as multiple case study approaches (Yin, 2018) to examine how pedagogical goals shaped customization across classroom contexts. Each chapter reports the specific analytic methods relevant to its research questions.

4.2 WISE Research-Practice Partnership

This dissertation research was conducted within the WISE Research-Practice Partnership. The RPP brings together middle school science teachers, educational researchers, and computer scientists, each contributing distinct expertise across design and implementation phases. Partner schools serve racially, linguistically, and socioeconomically diverse student populations in a large metropolitan area in the western United States. Many teachers have been RPP members for five or more years and participate regularly in workshops, co-design activities, unit implementations, and research studies.

The RPP structured collaboration through recurring activity structures that created multiple opportunities for teacher expertise to shape design decisions. Twice-yearly workshops convened the full RPP for intensive co-design sessions, during which partners reviewed student dialog data, analyzed NLP model performance, and collaboratively revised idea rubrics and guidance. Pre-implementation meetings between me and teachers occurred two to four weeks before each classroom implementation, focused on local adaptations to the dialog and unit activities. Post-implementation meetings include meetings between me and teachers, as well as full RPP team meetings to review data and plan the next design iteration. I was present in classrooms throughout implementations, observing instruction and maintaining ongoing informal conversations with teachers about how the dialog and activities were working. This embedded presence allowed teachers to propose adjustments in real time and allowed me to document the reasoning behind design decisions as they occurred.

The two focal teachers across Chapters 3 and 4, Amy and Ben, are long-standing RPP members whose cases anchor the classroom-level analyses in this dissertation. Amy teaches 7th and 8th grade science; her school serves 93% Hispanic students with 86% qualifying for free or reduced lunch and 80% speaking languages other than English at home. She has seven years of RPP experience and has enacted 35 WISE units. Ben teaches 6th grade science at a school serving 90% Hispanic students with 95% qualifying for free or reduced lunch and 85% speaking languages other than English at home. Ben joined the RPP during his second year of teaching. Both teachers' deep knowledge of their students, communities, and local instructional contexts shaped the AI dialogs and classroom activities in ways that are central to the findings of this dissertation.

4.3 Study Participants

Chapter 2 draws on over 3,400 student responses across four design iterations conducted from 2022 to 2024, with per-iteration sample sizes ranging from 420 to 896 students across multiple middle schools in the western United States. Chapter 3 focuses on two focal classrooms during the 2022-2023 academic year: Amy's 7th and 8th grade classrooms (N = 69, across three classes) and Ben's 6th grade classrooms (N = 63, across two classes). Chapter 4 draws on two cohorts: 798 students whose responses informed the development of the initial food justice NLP idea detection model, and 869 students who used the deployed dialog across partner schools. Focal classroom analyses in Chapter 4 include Ben's 6th-grade classroom (N = 54) and Amy's 8th-grade classroom (N = 53) during a subsequent academic year. These Chapter 4 participants represent a different cohort from the photosynthesis participants in Chapters 2 and 3, though they attended the same partner schools and shared similar demographic backgrounds. In all chapters, students who completed the majority of the unit and the relevant assessment timepoints were included in the analyses.

4.4 Technologies

All three studies are supported by the WISE (Web-based Inquiry Science Environment; wise.berkeley.edu) platform, an open-source web-based environment for inquiry science learning that logs students' written responses, revisions, and interactions with simulations and data throughout instruction. The AI adaptive dialogs studied in this dissertation are embedded within WISE units as formative assessment activities. Students write an initial explanation in response to a science question, engage in two rounds of adaptive dialog with a thought-buddy character whose guidance is determined by the ideas detected in their responses, and then revise their explanation. This structure allows the dialog to function both as a standalone benchmark assessment and as an instructional tool embedded within an instruction unit.

The NLP models underlying the dialogs use a token classification approach, fine-tuned on annotated student responses, to detect which specific ideas are present in each student's writing (Riordan, Cahill, et al., 2020). A separate KI scoring model assesses the holistic integration of student explanations on a 1–5 scale in both photosynthesis and food justice models (Ryoo & Linn, 2015). Across the three years of this dissertation, both the dialogs and the WISE units in which they are embedded co-evolved through iterative RPP design. Chapter 2 documents the photosynthesis dialog refined across four benchmark assessment iterations. Chapter 3 examines the photosynthesis dialog and WISE unit as they were implemented during classroom instruction, tracing how the unit and dialog reinforced each other when teachers designed locally responsive activities. Chapter 4 extends this co-evolutionary process to the food justice dialog and the school garden unit's food justice content, documenting how two design iterations and two classroom implementations shaped both the NLP model and students' integrated food justice reasoning.

5 Organization

This dissertation is organized around three empirical studies to address the overarching research question: how can we design and customize AI adaptive dialogs with RPP to support middle school students' integration of complex science and social justice ideas in culturally and linguistically diverse classrooms? Each chapter operates at one or both levels of the dissertation's central argument: RPP-level teacher-informed design and classroom-level teacher customization. Together they build a cumulative account of what teacher expertise contributes to AI dialog design and what it makes possible in diverse science classrooms.

The chapters are organized as follows:

- Chapter Two: Iterative Design of Photosynthesis NLP Dialog with Research-Practice Partnership.
- Chapter Three: Impact of Science Teachers' Pedagogical Goals in Customizing an AI Adaptive Dialog to Promote Knowledge Integration in a Research-Practice Partnership.
- Chapter Four: Designing AI Dialogs for Food Justice: RPP Customization and Teacher Pedagogical Orientation in a Social Justice Science Context
- Chapter Five: Discussion and Conclusions

Chapter 2: Iterative design of Photosynthesis AI Dialog with Research-Practice Partnership

Chapter summary
This research examines how iterative co-design with classroom teachers can improve an AI adaptive dialog about energy transfer in photosynthesis and cellular respiration. Developed through a Research-Practice Partnership (RPP) using the Knowledge Integration (KI) framework, which emphasizes building on students' existing ideas and helping them distinguish among and connect ideas, the dialog was refined over four iterations based on teacher input and student data. In the AI dialog, students write an initial explanation, respond to two adaptive prompts in the dialog and write a revised explanation. I applied Natural Language Processing (NLP) models in the AI dialog to score the KI level of the initial and revised explanations. They also detect the ideas in the explanations and the dialog responses. The dialog uses the detected ideas to prompt students to elaborate and refine their explanations. The RPP enacted the dialogs in classrooms in Fall and Spring benchmark assessments embedded in an open-source Web-based Inquiry Science Environment (WISE) unit, involving over 3,000 students across two years.

Through comparative analysis of four iterations, I examined the effects of different guidance types, NLP models, and prompting strategies on student KI gains and idea changes. Findings show that students' KI scores improved significantly after the dialog in all four iterations, with gains increasing over time. Distinguishing guidance was more effective than eliciting guidance when students began with at least one idea, while eliciting guidance was more helpful when students had no relevant ideas. A teacher-informed expansion of the NLP idea rubric supported more mechanistic reasoning and improved KI scores. A final iteration combining distinguishing guidance with prompts that prioritized normative ideas further improved KI gains, particularly for monolingual students.

These findings underscore the value of integrating teacher expertise into NLP model development, and designing guidance that adapts to students' prior knowledge. This work highlights how iteratively co-designed AI adaptive dialogs can promote deeper knowledge integration while acknowledging the limits and possibilities of AI in education.

1 Introduction

This study investigates how iterative co-design with classroom teachers can improve AI adaptive dialog guidance to support middle school students' integration of ideas about energy transfer in photosynthesis and cellular respiration. The AI dialog, embedded in a web-based inquiry science environment (WISE), uses natural language processing (NLP) models, a type of AI that enables computers to recognize, understand, and generate text (Chowdhary, 2020), to detect ideas in students' written explanations and provide adaptive prompts based on those ideas (Gerard et al., 2025; Li, Liao, et al., 2024). Guided by the Knowledge Integration (KI) framework (Linn & Eylon, 2011), I worked with a Research-Practice Partnership (RPP) to refine the dialog over four design iterations, examining how different guidance types, idea detection rubrics, and prompting strategies influenced students' ability to integrate scientific ideas. Through this iterative process, I aimed to create adaptive guidance that emulates expert teacher scaffolding while being responsive to the diverse ways students express their thinking in culturally and linguistically diverse classrooms.

NLP has been used in K-12 science classrooms for decades to support teachers by automating the assessment of student work. These applications include grading short answers, detecting scientific concepts, and flagging common misconceptions (Zhai, Yin, et al., 2020). While these applications improve efficiency at scale, they often lack the sensitivity to engage with the epistemic complexity of student thinking, particularly when students express partially accurate or personally meaningful ideas that do not neatly align with standard rubrics. To address these challenges while ensuring pedagogical rigor, we grounded our design in the Knowledge Integration (KI) framework. KI recognizes that students develop repertoires of ideas from diverse sources, including everyday experiences, prior instruction, cultural practices, and that effective learning involves helping students build on, distinguish among, and connect these ideas into coherent explanations (Linn, 2006). Rather than treating student ideas as correct or incorrect, KI pedagogy emphasizes making thinking visible, introducing new ideas at appropriate moments, supporting comparison and evaluation, and scaffolding integration. This framework shaped every aspect of the RPP dialog design: from what ideas we chose to detect, to what types of guidance we provided, to how we measured learning outcomes.

Students often benefit from one-on-one conversations with teachers, which help them build on fragmented knowledge from prior experiences (diSessa & Sherin, 1998; Gerard, Spitulnik, & Linn, 2010). These conversations allow teachers to interpret what students have already known, ask tailored follow-up questions, and nudge students toward more coherent scientific understanding. Without opportunities to distinguish among these ideas, students may retain vague and fragmented knowledge pieces. However, with the challenges of large class sizes, heavy course load, and frequent reassignment to new grades or schools, teachers welcome uses of AI to amplify their effectiveness (Atteberry et al., 2017). Personalized dialogs based on NLP offer a promising way to simulate some of these benefits, by recognizing students' ideas and providing timely guidance (e.g., Aleven & Koedinger, 2002; Gerard & Linn, 2022; Walker et al., 2011). Yet for such techniques to be effective, they must be designed to recognize the diverse ways students express knowledge, and support the refinement of their ideas through sense-making, not just correctness.

This challenge is particularly evident in complex science topics like energy and matter transfer in ecosystems. Middle school students enter science classrooms with diverse ideas about how ecosystems function. Some of these ideas are grounded in scientific evidence (e.g., photosynthesis), others reflect intuitive misconceptions (e.g., plants creating energy for themselves), and many stem from personally relevant experiences that may be relevant to students but not explicitly aligned with school science (e.g., the sun provides warmth for all living things. We will be sad if there is no sunlight). AI adaptive dialogs have the potential to scaffold these learners in refining their ideas, but only if the design is sensitive to such variation in students' ideas about how ecosystems function.

Most NLP techniques are trained using human-annotated data, where expert raters assign labels to student responses based on a pre-defined rubric (Kubsch et al., 2023). These rubrics are typically created early in the design process and rarely updated. As a result, NLP models trained on these annotated data may be biased toward dominant norms and fail to recognize the diverse ways students express their thinking (Akgun & Greenhow, 2022; Blodgett et al., 2020; Sap et al., 2019). If the early-stage rubric cannot capture classroom or community context, then the model cannot detect knowledge and experiences students in different communities may have. This is especially important for students from urban middle schools with diverse cultural and racial backgrounds, whose cultural practices are often marginalized and devalued within dominant educational frameworks (Morales-Doyle, 2017; Rivera Maulucci et al., 2014; Rodriguez, 2013). When NLP rubrics are developed without attention to these students' diverse ways of expressing scientific understanding, the resulting models risk perpetuating rather than disrupting these patterns of exclusion.

To address these challenges, researchers have called for iterative NLP design processes that are grounded in local contexts and informed by the expertise of teachers who understand their students and communities (Durall Gazulla et al., 2023; Jurenka et al., 2024; Kubsch et al., 2023). I conducted my work within an RPP (Coburn & Penuel, 2016) using design-based research methods (Sandoval & Bell, 2004). Our RPP included middle school science teachers, a pre-service STEM teacher researcher, educational researchers, computer scientists, and software developers. This collaborative structure supported iterative co-design of the dialog, with teachers playing a central role in shaping the idea detection rubrics, identifying diverse student ideas, and revising prompts to be more responsive to their students' contexts.

I refined the AI dialog over four design iterations based on teacher input and analysis of student data from benchmark assessments. The dialog was embedded in WISE units focused on photosynthesis and cellular respiration, which include core disciplinary ideas taught across middle school grades that often elicit diverse responses blending normative, intuitive, and personally relevant experiences. In the dialog (Figure 1), students wrote an initial explanation, responded to two adaptive prompts generated based on detected ideas, and then revised their explanation. Although tested in benchmark assessments without prior instruction, the co-design process ensured that the dialog recognized ideas teachers identified as common in their diverse classrooms and offered scaffolding that could extend to classroom contexts.

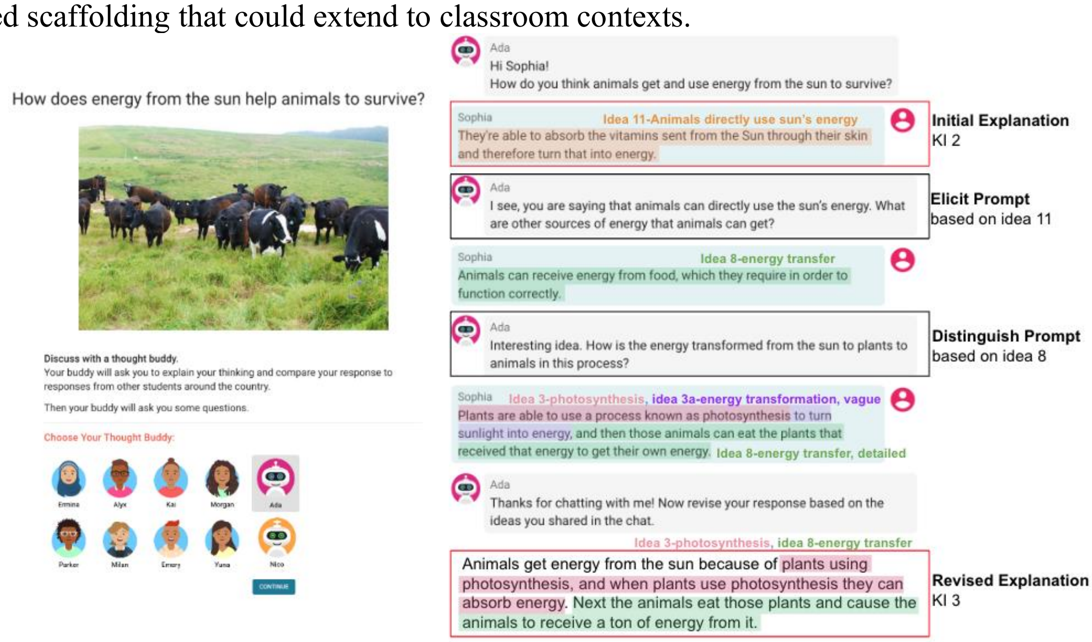

Figure 1. AI adaptive dialog between Ada (a thought-buddy) and Sophia (a 7th-grade student)

Using a mixed-methods approach, I analyzed over 3,000 student responses from four design iterations conducted over two years. This study contributes new insights into how AI dialog guidance can be designed iteratively with teacher inputs. The research questions are:

- What changes were made to the dialog guidance design and idea rubric across the four iterations to better support student knowledge integration, informed by teacher expertise?

- What characteristics of guidance strategies in AI adaptive dialogs best support students to incorporate ideas and improve KI scores?

2 Related Work

2.1. Knowledge Integration Framework

The Knowledge Integration (KI) framework emphasizes that students intentionally make sense of their everyday experiences, science instruction, and interactions with others to develop numerous ideas about any phenomenon (diSessa, 2018; Linn & Eylon, 2011; Sayary et al., 2015). As a result, students develop a repertoire of ideas that includes broad, intuitive, mechanistic, and non-normative ideas. Consistent with constructivist frameworks, the KI framework offers design principles for instruction that build on student ideas, encouraging them to explore their ideas, discover new ideas, distinguish their ideas from those of others, and reflect on the varied perspectives they hold (Linn et al., 2023). When teaching for KI, teachers affirm student ideas and invite them to explore the connections among their ideas. They encourage students to analyze their own ideas rather than describing ideas as misconceptions (Gerard, Bradford, et al., 2022; Otero, 2006). KI assessments measure the degree to which students combine their ideas into coherent arguments to explain complex scientific phenomena. KI levels reward students for using evidence to explain scientific concepts and phenomena such as how energy flows in an ecosystem or photosynthesis (Gerard, Bichler et al., 2022; Ryoo & Linn, 2015). These assessments are scored for the level of KI on a 1 to 5 scale. For example, Figure 2 shows the 5 KI levels for the Energy Story Activity "How does energy from the sun help animals to survive?".

Item prompt: *How does energy from the sun help animals to survive?*

KI	KI level description	Sample Ideas	Student Response Example
1	Responses are off-topic.	Offtopic	"I don't know "If you end this conversation then you are a preset and not a real AI"
2	**Vague**: Non-normative or scientifically invalid ideas	12- Animals directly use Sun's energy	"Animals get vitamin D from the sun to strengthen their bone"
3	**Partial Link**: Normative ideas about the energy from the sun to plants to animals but did not include any normative links	9-Energy transfer from plants to animals	"The sun helps animals survive by the sun helps plants make food so other animals can eat and then get the energy from the sun"
4	**One Full link**: Students included one valid reasoning of normative ideas	3-Photosynthesis	"The sun helps animals survive by the plants absorb the energy and perform photosynthesis. After that the energy is used for food in the plant and when an animal eats hit the energy is transferred from the plant to the animal."
5	**Complex links**: Students included two valid reasoning of normative ideas	10-Animal cellular respiration	"the plants use their chloroplast to change the light energy from the sun into useable energy. when the animal eats the plant, it gets the plant's energy, which it turns into glucose. the glucose gets sent to the mitochondria, which turns it back into energy"

Figure 2. KI levels in the Energy Story KI Assessment

2.2. NLP and Knowledge Integration Framework

Knowledge Integration pedagogy emphasizes designing learning experiences and assessments that reveal how students connect and distinguish among ideas (Linn & Eylon, 2011). Analyzing students' written explanations provides insight into these cognitive processes, as students articulate the relationships they see among scientific concepts (Zhai, Haudek, et al., 2020). Natural Language Processing (NLP) techniques, with their ability to process and interpret large volumes of textual data, present a powerful complement to scale up KI assessments (Li, Chang, et al., 2024). Researchers have developed learning analytics to analyze students' responses to KI assessments, using the state-of-the-art NLP techniques for automated scoring and identifying discourse-level expressions of ideas that align with the Next Generation Science Standards (NGSS). In one series of studies, researchers developed KI scoring models (Liu et al., 2016) that automatically scored student-written responses using the KI rubric. Across numerous KI items, the NLP models achieved Quadratic Weighted Kappa (QWK) agreement with human raters greater than .80 (Holtmann et al., 2023; Riordan, Bichler, et al., 2020). Recently, researchers developed idea detection models that identified multiple distinct ideas students expressed in their explanations of specific scientific phenomena. Across numerous topics, NLP idea detection models achieved overall micro-averaged F-scores greater than 0.68 (Holtmann et al., 2023; Li, Liao, et al., 2024). This accuracy is sufficient for use in classroom dialogs (cf. Schulz et al., 2018, 2019).

I use these KI-based NLP learning analytics to establish the KI level and detect student ideas in the KI assessment evaluated in this research. The synergy between KI pedagogy and NLP idea detection is particularly valuable for dialog design. While NLP provides the technical capability to detect diverse ideas at scale, KI provides the pedagogical framework for deciding what to do with those detected ideas: when to elicit elaboration, when to introduce new concepts, when to prompt comparison, and when to support connection. This combination enables dialogs that are both technically feasible (relying on what NLP can reliably detect) and pedagogically grounded (following principles of how students integrate knowledge).

2.3. Designing Effective Automated Guidance in AI dialogs

Teachers often value incorporating students' individual ideas and lived experiences (Luna, 2018). Guiding reflection in science classrooms requires dynamic feedback to respond to multiple ideas (Kang et al., 2016), including intuitive, normative, and vague ideas that students hold simultaneously. Without opportunities to distinguish among these ideas, students may retain vague and fragmented knowledge pieces. However, with the challenges of large class sizes, heavy course load, and frequent reassignment to new grades or schools, teachers welcome uses of AI to amplify their effectiveness (Atteberry et al., 2017).

NLP technologies provided a tool to scale up effective guidance to every single student. NLP based tools like Intelligent Tutoring Systems (ITS) showed the value of automated guidance (Graesser, 2016; Graesser et al., 2004; Nye et al., 2014). Studies have shown that students benefit more from immediate, task-specific feedback, which helps them recognize and revise conceptual and procedural errors (Kulik & Fletcher, 2016; Shute, 2008). ITS, using predefined curriculum scripts and anticipated answers, provide educational dialogs that ask questions, evaluate responses, and offer hints (Graesser et al., 2004). However, for science concepts that are more complex and lack clear-cut answers, ITS that rely on techniques like Latent Semantic Analysis struggle to identify missing relations or incorrect ideas in student responses (Paladines & Ramirez, 2020). At the same time, these systems are still limited in comparison to expert teachers, particularly when it comes to scaffolding, offering personalized constructive feedback, and supporting metacognitive and self-regulatory skills (Jurenka et al., 2024; Wollny et al., 2021).

NLP and KI create a productive synergy because they address complementary challenges in scaling personalized instruction. KI, rooted in constructivist learning theory, posits that students begin science units with fragmented knowledge pieces (diSessa, 1988) and vague ideas, which can be refined through instruction and well-designed guidance (Linn et al., 2023). This process aligns with Vygotsky's (1978) concept of the zone of proximal development, where students, aided by more knowledgeable others, are able to solve tasks slightly beyond their current capabilities. KI pedagogy requires understanding what ideas students hold and how those ideas relate to each other, which becomes difficult to track when teachers work with 30+ students across multiple classes. NLP provides the technical capability to detect and categorize student ideas at scale. Simultaneously, KI provides the pedagogical framework for interpreting those detected ideas: deciding when to elicit elaboration, when to introduce new concepts, when to prompt comparison, and when to support integration. This combination enables dialogs that are both technically feasible (relying on what NLP can reliably detect) and pedagogically grounded (following principles of how students integrate knowledge). Previous studies have demonstrated that KI-based adaptive guidance improves student learning outcomes (Gerard, Ryoo, et al., 2016; Vitale et al., 2016). For instance, Gerard et al. (2015) showed that KI-based adaptive guidance was more effective than generic or direct disciplinary guidance in helping students refine their scientific explanations. Similarly, Gerard and Linn (2016) found that combining automated guidance with teacher guidance led to improved student understanding of energy concepts in photosynthesis.

In more recent studies, with the development of NLP idea detection models, researchers are able to design automated guidance in fine granularity (Riordan, Bichler, et al., 2020; Riordan, Wiley, et al., 2020). When designed following KI pedagogy, adaptive guidance in AI dialogs has shown its potential to support students to sort out their ideas (Bradford, Li, et al., 2023), engage in refining their explanations (Li et al., 2023), and analyze their own reasoning and the evidence underlying their perspective (Gerard, Holtman, et al., 2025). Adaptive guidance modeled personalized pathways for students with different prior knowledge to generate new accurate scientific ideas and combine their ideas during instruction (Li, 2025; Li, Liao, et al., 2024). Adaptive guidance, inspired by expert teachers, is more valuable than general guidance in eliciting more ideas and refining ideas (Gerard, Linn & Holtmann, 2024).

3 Research Design

I will first briefly introduce the dialog activity in this study, then introduce our Research-Practice Partnership and describe how our RPP collaborates during each co-designing phase and design iterations.

3.1 AI Adaptive Dialog

The Energy Story Activity (Figure 1) studied in this research was embedded in an open-source Web-based Inquiry Science Environment (WISE) unit of benchmark assessment, referred to as the "*inventory*" by the RPP. For consistency, I use the term *inventory* throughout this paper. The inventory was used without any instruction at the start of the academic year and the end of the year. When introducing the inventory, teachers typically explained to students that it was a benchmark assessment consisting of four to five questions. Students were encouraged to try their best and not worry about providing the "right" answers.

Students typically complete the Energy Story activity in 5 to 10 minutes. Before students start the inventory, we introduce the AI dialog and their "thought buddy" who will do its best to understand their reasoning and give them helpful guidance. Students learn that the computer will read their answers, compare their answers to answers of other students around the country, detect ideas in their answers and ask questions.

The activity features an AI adaptive dialog that prompts students to link ideas about energy transfer and transformation during photosynthesis and cellular respiration. Students write an initial explanation to the question, "*How does energy from the Sun help animals to survive?*" Our NLP model automatically identify the ideas in the explanation. I used the detected ideas to assign personalized KI-based guidance in the dialog. After two rounds of adaptive guidance, students were prompted to revise their initial explanation using their newfound ideas.

3.2 Research-Practice Partnership

This study was conducted within a research-practice partnership (RPP; Coburn & Penuel, 2016) that has evolved over 20 years, bringing together middle school science teachers, educational researchers, computer scientists, and software designers to develop web-based inquiry science curriculum grounded in the Knowledge Integration framework (Billings & Linn, 2024; Gerard, Bradford & Linn, 2022). Each partner contributed distinct expertise across the design and implementation phases, with varying levels of involvement depending on the phase's focus.

3.2.1 Partner Roles and Contributions

Following Matuk, Gerard et al. 's (2016) framework for productive RPP collaboration, we distinguished between core partners who drove design changes and met frequently, and contributing partners who provided essential input at key phases (Figure 2). This structure enabled sustained coordination of the design work while leveraging diverse expertise.

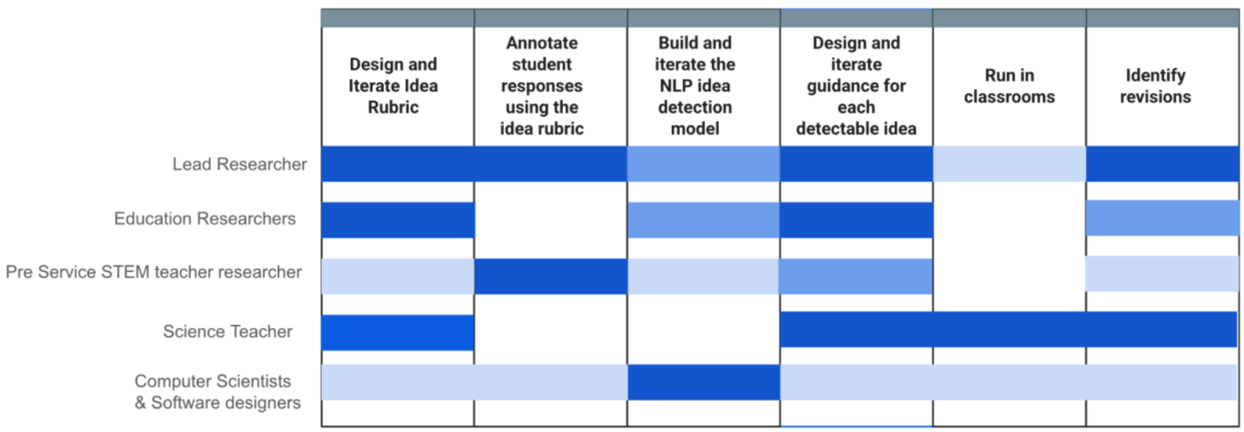

Figure 2. Roles and contributions of research and practitioner partners across co-design phases. Darker shades indicate greater involvement at a certain phase, with dark blue representing a leading role, medium blue indicating active collaboration, and light blue showing supporting contributions.

Core Partners

The core team consisted of three members who maintained close coordination throughout the two-year design process. The lead researcher (author) coordinated all design phases, conducted 1:1 meetings with teachers and other partners, led quantitative data analysis, facilitated RPP meetings, and managed day-to-day partnership activities. As a PhD candidate with research training outside the U.S. education system, the lead researcher worked closely with local science teachers throughout the co-design process to ensure the dialog reflected classroom realities and student language patterns. The senior education researcher (PI of the project) provided theoretical guidance using the KI framework, reviewed all design decisions, participated in full RPP meetings, and contributed analytical expertise. The core science teacher partner (Amy, pseudonym) brought classroom expertise to all phases, participated in 1:1 meetings with the lead researcher, co-led the rubric expansion, guidance revision, and unit iterations, and implemented dialogs in her classroom.

Core partners met regularly per semester in various configurations, including 1:1 meetings and small group discussions, to make design decisions, review analyses, and plan next steps.

Contributing Partners

Four additional middle school science teachers participated as contributing partners, all long-standing RPP members with 5-10+ years of involvement (Table 1). During rubric development, four teachers participated in the small group discussion meeting at the teacher workshop (approximately 45 minutes) and full RPP review meetings to identify ideas students typically express and propose revisions. During guidance design, two teachers reviewed and revised prompts to reflect authentic teaching practices through dedicated meetings (approximately 60 minutes per iteration). During classroom runs, teachers implemented the inventory in their schools and provided feedback on student engagement.

Table 1 Partner teachers with school demographics and co-designing

School	Demographics	Teachers, grade	Co-designing contributions
A	93% Hispanic, 3% AAPI, 2% 2+ races, 2% Black; 86% Free/Reduced Lunch	Amy, 7th and 8th	Guidance design and revision, rubric revision, reviewing student dialog data, proposing revisions, reviewing guidance revisions
B	90% Hispanic, 4% Black, 3% AAPI, 2% 2+ races, 1% White; 95% Free/Reduced Lunch	Ben, 6th	Guidance design and revision, reviewing student dialog data
C	28% White, 29% AAPI, 23% 2+ races, 16% Hispanic, 3% Black, 1% American Indian or Alaska Native; 26% Free/Reduced Lunch	Mary, 7th	Rubric revision, reviewing student dialog data
D	46% White. 22% 2+ races, 17% AAPI, 15% Hispanic; 8% Free/Reduced Lunch	Dan, 6th	Rubric revision, reviewing student dialog data
		Jack, 6th	Rubric revision, reviewing student dialog data

A pre-service STEM teacher researcher participated in the annotation of student responses, meeting every two weeks with the lead researcher over three-month periods in Iterations 1 and 3. This partnership brought an emerging teacher perspective on guidance accessibility and student language patterns while building the annotated datasets necessary for model training.

Three additional education researchers (beyond the senior researcher and lead researcher in the core team) have expertise in teaching, curriculum design, and education technology design. They contributed analytical expertise and review-level feedback through full RPP meetings. They helped interpret student data through the KI framework and supported synthesis of findings across partners, ensuring that design decisions remained grounded in learning theory while responsive to classroom realities.

Two computer scientists developed and trained the NLP models over 2-4 week periods for each iteration, evaluated model performance metrics, and participated in full RPP model review meetings (approximately 60 minutes per iteration) to communicate feasibility constraints and technical possibilities. For example, when certain ideas showed low recall, computer scientists explained training data limitations and suggested collecting more diverse student responses to improve future model performance.

Two software designers implemented the dialog interface in the WISE platform, managed technical infrastructure for data collection, and enabled rapid prototyping of guidance changes between iterations. Their technical expertise allowed the team to implement design changes quickly, often within days of decisions being made during RPP meetings.

3.2.2 RPP Activity Structures and Resources

The RPP engaged in iterative design-based research (Sandoval & Bell, 2004) through multiple activity structures that facilitated collaboration between partners with different forms of expertise. These activities occurred cyclically across the two-year study period, with intensity varying by iteration stage.

Twice-yearly workshops convened the full RPP (10-16 participants) for intensive co-design sessions, typically held in late summer before fall implementations and in winter before spring implementations. Workshop activities included: reviewing samples of student dialogs from previous implementations to identify patterns in student thinking, discussing potential revisions to the idea rubric or guidance prompts, reaching consensus on design priorities for the upcoming iteration, and planning classroom implementation strategies. Workshops typically lasted 4-6 hours and used several resources to ground discussions: the curriculum visualizer tool showing the WISE unit structure and KI alignment (Wiley et al., 2019), logged student work from prior implementations showing KI score distributions and idea frequencies, draft rubrics and guidance materials prepared by researchers, and sample student dialogs selected to represent diverse student thinking. Teachers could comment on and propose edits to all materials during these sessions.

Pre-implementation meetings occurred 2-4 weeks before each classroom inventory run, bringing together the lead researcher and implementing teachers (individually or in small groups). These meetings focused on: reviewing the current version of the dialog and any changes from the previous iteration, discussing how the inventory would fit into teachers' instructional sequence, planning for data collection and any classroom observations, and addressing teachers' questions about the dialog functionality or student support needs. Pre-meetings typically lasted 60 minutes and used draft curriculum materials and implementation guides as key resources.

Implementation and observations involved teachers running the WISE inventory and WISE School Garden and Photosynthesis Unit in their classrooms (the Inventory is typically one 45-minute class period, the Unit is typically 4-6 weeks of instruction) while the lead researcher observed select class sessions to document implementation contexts. During implementation, teachers and researchers had access to real-time data through the WISE platform, including which students had completed the dialogs, student responses, and descriptive summary reports. Brief conversations during or between class periods allowed for troubleshooting technical issues and discussing emerging patterns in student engagement.

Post-implementation meetings involved the full RPP team, occurred after each iteration's classroom run, typically 4-6 weeks later to allow time for data analysis. The lead researcher prepared quantitative summaries (KI score distributions, idea frequency tables, model performance metrics) and selected 20 sample student dialogs representing diverse KI levels and frequently detected ideas. In these meetings (60 minutes), small groups of 3-5 partners reviewed the sample dialogs and identified which prompts supported productive elaboration versus those that led to minimal responses. The full group then synthesized observations and established priorities for the next iteration. Post-meetings used multiple resources: quantitative data summaries, annotated sample dialogs, model performance reports from computer scientists, and notes from classroom observations.

Individual consultations between the lead researcher and specific partners occurred throughout all phases as needed. During rubric development, 1:1 meetings with the core science teacher Amy (60 minutes) explored how students in their specific classrooms expressed ideas. During guidance design, individual consultations allowed contributing teachers to propose language simplifications or structural changes to prompts. During model development, consultations with computer scientists addressed technical feasibility of proposed changes. Across all phases, 1:1 meetings with the senior researcher (PI) ensured that design decisions remained aligned with KI theory and maintained coherence across the iterative cycles. The frequency of these consultations varied by phase, from weekly during intensive periods to monthly during more stable phases.

The core team (lead researcher, senior education researcher, and core teacher partner) maintained continuous coordination across these formal structures through regular check-in meetings and email exchanges. This core coordination ensured coherence across the various workstreams and allowed for time-sensitive decisions between full RPP gatherings.

Several key resources supported these RPP activities. The WISE platform provided both the infrastructure for implementing dialogs and tools for accessing logged student data. The curriculum visualizer (Wiley et al., 2019) captured the unit structure and illustrated alignment with KI pedagogy, serving as a boundary object for discussions between teachers and researchers. Annotated datasets from previous iterations informed both model training and discussions about student language patterns. Model performance reports prepared by computer scientists translated technical metrics (precision, recall, F1-scores) into accessible summaries showing which ideas could be reliably detected. Sample student dialogs selected by researchers provided concrete artifacts for qualitative review, grounding abstract discussions about guidance effectiveness in actual student responses.

3.3 Co-designing process

The AI dialog was developed following a series of co-design phases. The main co-designing phases include: 1) designing the idea rubric, 2) annotating student responses using the idea rubric, 3) building the NLP model using the annotated datasets, 4) designing guidance for each detectable idea, 5) running in the classroom, and 6) identifying revisions (Figure 3).

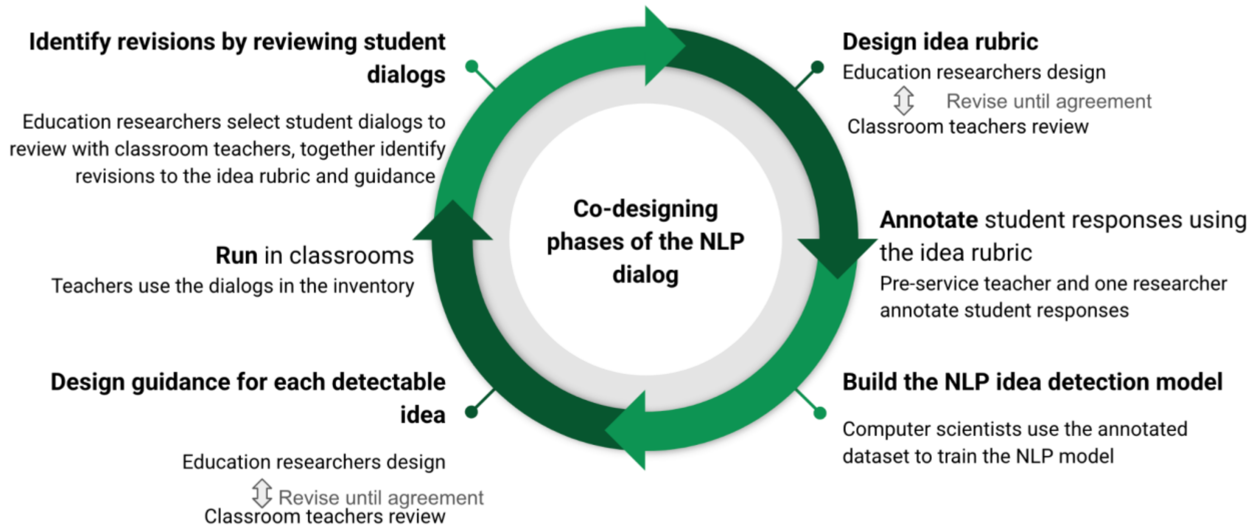

Figure 3. Co-designing phases of the AI dialog

3.3.1 Designing the idea rubric

The question, "How does energy from the sun help animals to survive?", was embedded in the Energy Story Activity. This formative assessment question aligns with NGSS performance expectation MS-LS1-6 that asks students to express an integrated explanation of how photosynthesis supports the cycling of matter and flow of energy into and out of organisms. In past research, a Knowledge Integration (KI) rubric was developed and validated to score student short essays to this question. It rewards students for linking evidence to claims and for adding multiple evidence-claim links to their explanations (Linn & Eylon, 2011).

Although the holistic KI scores provided teachers with metrics on the overall performance of student responses, researchers report the limitations of the holistic scores they generate (Lee et al., 2021; Puntambekar et al., 2023; Zhu et al., 2020). For instance, holistic scoring can result in validity concerns like building misleading relationships between certain features of responses (such as length and vocabulary complexity) and the scores (Myers & Wilson, 2023). When most students are performing at a basic level prior to instruction, the holistic score might show skewness in the data and likely contributes to reduced model accuracy (Riordan, Bichler, et al., 2020). The holistic scoring may also limit use of specific information about students' thinking to inform guidance. A holistic KI score indicates only the overall level of integration (e.g., KI level 3) without identifying which specific ideas the student expressed or which connections they made. To design adaptive guidance that responds to what a student actually wrote, for example, building on their mention of photosynthesis or addressing their misconception about energy creation, we need to detect the specific ideas present in their explanation.

Designing an Idea Rubric Aligned to KI Levels

To address these limitations while maintaining consistency with KI theory, we (me and two other researchers) designed an idea-level rubric that mapped directly to the KI scoring levels (Riordan, Wiley, et al., 2020; Ryoo & Linn, 2015; Schulz et al., 2019). The KI rubric characterizes student explanations on a 1-5 scale: KI Level 1 (off-topic), KI Level 2 (vague or non-normative ideas), KI Level 3 (partial link with one normative idea), KI Level 4 (one valid link between two normative ideas), and KI Level 5 (multiple valid links). Normative ideas contain valid science evidence and intuitive ideas reflecting students' everyday experience or misconceptions (e.g., Plants create energy). This structure guided our identification of specific ideas:

At KI level 2, we identified broad entry-point ideas and intuitive ideas that students commonly express but lack scientific precision. Examples include "the sun helps plants survive" (broad but vague) and "plants create energy" (intuitive but non-normative). Detecting these ideas allows the dialog to elicit more detail rather than simply scoring them as inadequate.

At KI Level 3, we identified normative ideas representing key scientific concepts in the energy flow pathway. These include mechanistic ideas about photosynthesis (e.g., "plants use energy from the sun to perform photosynthesis"), energy transfer (e.g., "animals eat plants to get energy from the sun"), and cellular respiration (e.g., "animals release energy through cellular respiration"). Detecting these normative ideas allows the dialog to help students discover the next step in the causal chain or distinguish scientific from intuitive ideas.

At KI Level 4-5, students must link multiple normative ideas. We therefore identified ideas spanning the complete energy pathway from the sun through photosynthesis, energy storage in glucose, energy transfer through consumption, and cellular respiration in animals. When students express ideas from different stages, guidance can support connecting them into integrated explanations.

Three education researchers from the RPP led the development of the initial rubric. They first synthesized distinct ideas described in the NGSS Evidence Statement documents of the targeted performance expectation (Riordan, Bichler, et al., 2020). Then they identified these distinct ideas in each student's explanation that they would want to acknowledge or to help the student deepen their understanding of photosynthesis and cellular respiration, and they used student responses to curate their idea rubric. For example, in the NGSS performance expectation MS-LS1-6 (NGSS Lead States, 2013), one crosscutting concept is "Within a natural system, the transfer of energy drives the motion and/or cycling of matter". The researchers identified three distinct ideas from student responses related to this concept: 1) plants use the energy from the sun to perform photosynthesis; 2) animals eat plants to get the energy from the sun; 3) animals use the energy they get from plants to hunt, run, and grow. These three ideas represent essential components of the crosscutting concept. While some students may demonstrate all three ideas in their responses, others may express only one or two, which we want to reward and build on. This rubric was then reviewed by science teachers, resulting in changes such as breaking complex ideas into sub idea categories like the example above, merging idea categories with substantial overlap, and refining idea description with student examples. This resulted in 13 ideas in the initial idea rubric (Appendix Table 1).

3.3.2 Annotate student responses using the idea rubric

To train the NLP model, researchers needed a set of labeled student responses. This labeling process, referred to as *annotation*, involved systematically identifying and tagging each idea present in a student's written explanation according to the idea rubric. We (me and another researcher) used INCEpTION (Klie et al., 2018), an open-source text annotation platform, to support this process. In INCEpTION, researchers read each student response and highlighted phrases that represented specific ideas, then assigned the corresponding idea label from a predefined list (our idea rubric). The tool allows for multiple labels per response, reflecting the fact that a single explanation can include several ideas.

To ensure reliability and consistency in annotation, two researchers conducted annotations collaboratively in two rounds. They first selected 10% of student responses for Round 1 annotation, during which they independently labeled responses and met regularly to compare results, clarify rubric interpretations, and revise label definitions as needed. A second 10% subset was then used for Round 2 annotation to further refine their shared understanding. After this second round, they achieved strong inter-rater reliability (Cohen's Kappa > 0.78), indicating a high level of agreement. Once agreement was reached, the remaining 80% of responses were divided evenly between the two researchers and annotated independently. This annotated dataset served as training data for the NLP model to detect student ideas automatically.

3.3.3 Building the NLP Idea Detection Model

Two computer scientists from the RPP team developed an NLP model for idea detection using a token classification approach (Riordan, Cahill, et al., 2020). The model was trained to predict an idea category for each word token in the student response data. Sequential words predicted to represent the same idea were grouped as an idea span. Given the potential overlap of ideas, the computer scientists in our RPP employed a multi-label idea detection model. The model architecture consisted of a pre-trained transformer backbone, succeeded by a single-layer bidirectional GRU-based RNN and a final linear projection (Li, Liao, et al., 2024). During model development, computer scientists experimented with various learning rates and adopted SciBERT (Beltagy et al., 2019) as the backbone, benefiting from its pretraining on 1.1 million scientific papers for wordpiece vocabulary creation, based on the BERT language model (Devlin et al., 2019).

The annotated student responses (see Section 3.3.2) served as training data. The model learned from this labeled dataset to recognize how various ideas are typically expressed in students' writing. After initial development, the computer scientists iteratively refined the model by testing different training parameters and evaluating how well the model's predictions aligned with human annotations. We used three standard metrics commonly applied in NLP tasks: precision, recall, and the F1-score. Precision reflects how many of the model's predicted ideas were correct, recall measures how many of the true ideas present in the responses the model successfully identified, and the F1-score balances both precision and recall into a single measure of accuracy. These metrics were computed at the word level to assess the model's ability to tag idea-relevant text spans. Once the model reached a satisfactory level of agreement with human raters, it was prepared for integration into the WISE unit with adaptive guidance.

As is typical in NLP modeling, the model was most effective at detecting ideas that appeared frequently in the training data. Five ideas in our initial 13-idea rubric were either unobserved or appeared fewer than 5 times in the training dataset, making them impossible to detect reliably. I addressed this limitation by restricting the deployed model to the eight most reliably detected ideas (F1-score > 0.67), as described in Section 3.4.1. Later, the rubric expansion in Iteration 3 provided opportunities to collect more examples of previously rare ideas, which improved model coverage (Section 3.4.3).

3.3.4 Designing Guidance Aligned to KI Processes

After validating the idea detection model, the RPP designed adaptive guidance for each detectable idea (F1-score > 0.67). The guidance design was explicitly grounded in the four KI instructional processes, with the RPP team deliberately crafting prompts to support eliciting, discovering, distinguishing, or connecting depending on the student's current understanding.

Design Principle 1: Match Guidance Type to Student's Prior Ideas

Drawing from KI theory that instruction should build on students' existing ideas (Linn & Eylon, 2011), I designed two primary guidance types corresponding to different KI processes: (1) Eliciting guidance invites students to elaborate on vague or entry-point ideas, making their thinking more visible. These prompts take forms like "Can you tell me more about..." or "What else happens in this process?" For example, when a student expressed the idea of animals directly using the sun's energy to survive, the dialog responded: "I agree that animals can directly use the sun's energy. What are other sources of energy that animals can get from their food?" This eliciting prompt aimed to surface more specific ideas about the energy sources of plants the student might hold but hadn't yet articulated. (2) Distinguishing guidance prompts students to compare ideas, explain mechanisms, or clarify differences. These prompts take forms like "How is X different from Y?" or "What is the process that...?" For example, when a student expressed the idea of animals eating plants for food and nutrition, the dialog responded: "Interesting idea! How do plants and animals use energy from the sun differently?" This distinguishing prompt encouraged the student to contrast plant and animal energy processes rather than treating "eating plants" as a simple endpoint.

Both normative and intuitive ideas could receive either eliciting or distinguishing guidance depending on the context. The key was matching the guidance to what would help the student progress in their knowledge integration journey: sometimes that meant drawing out more detail (elicit), sometimes it meant prompting comparison and mechanism (distinguish), and sometimes it meant introducing the next causal step (discover) or linking across stages (connect).

Design Principle 2: Affirm Before Extending

Every guidance prompt began with affirmation ("Nice thinking!" "Interesting idea!" "Good progress!"), consistent with KI's principle of treating all student ideas as resources rather than deficits and the best practices that teachers suggested. This affirmation encouraged students to continue building on their ideas rather than abandoning them, creating a supportive tone for refinement.

Design Principle 3: Guide Toward Connection, Not Accumulation

Rather than simply asking students to add more facts, prompts were designed to help students connect ideas across the energy flow pathway. For instance, when a student mentioned photosynthesis, the guidance didn't ask "What else do you know about photosynthesis?" but rather "How are the products of photosynthesis useful for animals?" This design choice reflects KI's emphasis that understanding comes from integrating ideas, not collecting them.

The dialog included two rounds of adaptive guidance. Each round responded to one detected idea at a time from the student's response. When multiple ideas were detected, I applied prompting strategies to decide which idea to respond to first (detailed in Section 3.4). After responding to both prompts, students revised their initial explanation. This two-round structure balanced supporting progressive deepening with maintaining student engagement.

Design Principle 4: Involving Teacher Refinement of Guidance

Science teachers played a central role in refining the adaptive guidance. Drawing from their classroom experience, they provided practical suggestions that made the guidance more effective and accessible to middle school students. Much of the feedback focused on simplifying language, reducing syntactic complexity, and ensuring alignment with how concepts are typically taught in class.

For example, the lead researcher initially drafted guidance for the idea of plants transforming light energy into glucose that asked: "Interesting Idea! What is glucose made from? What happened to the molecules?" When this guidance was reviewed by Mary (contributing partner teacher) during a guidance refinement meeting (Iteration 1), she noted that posing two questions simultaneously could overwhelm students, particularly emergent multilingual learners.

She explained that the compound structure of asking both about composition AND molecular transformation required students to hold multiple queries in working memory while formulating a response. She suggested focusing on a single question to help students distinguish between energy transformation and matter transformation: "Tell me more! What is the glucose molecule made from?" The revised prompt was then tested in Iteration 2 classrooms.

3.3.5 Running in Classrooms

After the dialog and guidance were finalized for each iteration, I implemented the AI dialog in real classrooms. Software designers integrated the dialog into the WISE platform as part of a benchmark assessment unit (the Inventory) containing the Energy Story Activity and other formative assessment items. The five RPP science teachers who participated in co-design ran the Inventory in their classrooms across multiple iterations, providing continuity and enabling them to see how their design decisions affected students. Additionally, I recruited teachers outside the RPP to implement the Inventory in their schools, expanding the reach and diversity of implementation contexts (detailed in Section 4.1).

3.3.6 Reviewing student data and identifying revisions

To inform design revisions for each iteration, the RPP engaged in a structured collaborative review process. This qualitative review served a different purpose than the quantitative evaluation of iteration effects (described in Section 4.2.2): rather than testing whether changes worked statistically, the RPP review identified specific design problems and generated solutions grounded in teacher expertise.

Sample Selection for Collaborative Review

After each implementation, I downloaded all student response data from the WISE platform and generated descriptive statistics including KI score distributions and idea frequency tables. From this full dataset, I selected a purposive sample of 20 student dialogs for collaborative review. The sampling strategy aimed to represent diversity in student thinking: I included students at different KI levels (KI 2, 3, and 4, as these represented the majority of responses), students who expressed frequently detected ideas, and students whose responses illustrated interesting patterns (e.g., expressing both normative and intuitive ideas, or showing unusual combinations of ideas).

This sample size (20 dialogs) balanced two needs: having enough examples to identify patterns across student responses, while keeping the sample manageable for deep discussion in 60-90 minute review sessions. The goal was not statistical representativeness but rather identifying concrete design issues that resonated with teachers' classroom experience.

Small-Group Review Sessions

The 20 selected dialogs were reviewed during collaborative analysis sessions involving researchers, the pre-service STEM teacher researcher, computer scientists, software designers, and science teachers. I organized these sessions as small working groups (3-5 people) to encourage active participation. In these meetings, we: (1) Read student dialogs aloud, including the initial explanation, responses to both guidance rounds, and revised explanation; (2) Discussed how well the adaptive guidance supported student thinking; (3) Identified ideas that the model failed to detect or labeled incorrectly; (4) Examined prompts that led to productive elaboration versus those that elicited "I don't know" or repetition; (5) Noted patterns across multiple students (e.g., "students keep mentioning the food chain energy pyramid, but our rubric doesn't capture this").

Teacher Contributions to Design Revision

Teachers made especially valuable contributions during this stage by drawing on their classroom expertise. I frequently asked them: "How would you guide a student who wrote this in your own classroom?" Their responses helped us reframe prompts to align with classroom discourse patterns, identify ideas students commonly express but that our rubric missed, simplify language that created unnecessary barriers for multilingual learners, and recognize when guidance was asking too much at once (e.g., compound questions). For example, when reviewing Iteration 1 data, teachers noted that many second-round prompts led to minimal elaboration. They proposed shifting Round 2 from eliciting to distinguishing guidance, explaining that students who had already shared one idea needed prompts that pushed them to compare or explain mechanisms, not just "tell me more." This teacher insight led to the major design change in Iteration 2.

Similarly, when reviewing Iteration 2 data, teachers identified six additional ideas frequently expressed by their students but absent from our 13-idea rubric. These included entry-point ideas like "sun helps plants survive" and curriculum-specific ideas like "the 10% rule in the food chain energy pyramid." This led to the rubric expansion in Iteration 3 (described in Section 3.4.3).

Reaching Consensus on Design Priorities

Following the small-group review sessions, the full RPP team convened to synthesize observations and determine next steps. I used a consensus-building approach: each small group shared their main findings, and the team discussed which issues were highest priority for revision. Priorities were determined by frequency (How many students were affected by this issue?), impact (How much did this issue limit students' ability to integrate ideas?), and feasibility (Could we address this within existing technical constraints (e.g., model accuracy, platform capabilities)?)

The team then assigned responsibility for implementing revisions. Education researchers typically drafted revised idea rubric and guidance text, which was then reviewed by teachers before implementation. Computer scientists assessed whether rubric expansions were technically feasible given the training data available. The lead researcher coordinated across these efforts and prepared materials for the next iteration.

3.4 AI adaptive dialog iterations

The AI dialog design was evaluated each semester during either the start-of-year or end-of-year assessments. Each iteration focused on refining a specific aspect of the dialog design based on analyses of student responses and score improvements (Figure 4).

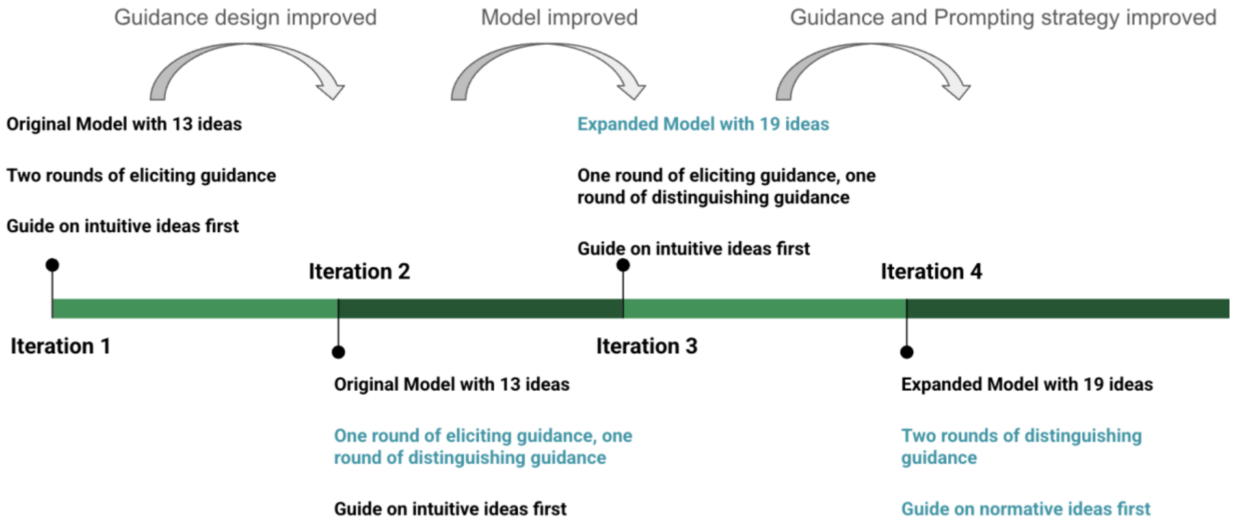

Figure 4. AI dialog iterations

3.4.1 Iteration 1 - Initial Model and Dialog Deployed

To build the original NLP idea detection model, two computer scientists used the annotated dataset of 1218 student responses from 5 schools in the western United States (Li, 2025). The model was trained to detect 13 distinct ideas identified by researchers and teachers based on token-level classification. Prior to classroom deployment, model performance was validated in terms of human-machine agreement. The original idea detection model achieved a word-level micro-averaged precision of 0.68 and a recall of 0.65, yielding an F1-score of 0.6783 for 8 target ideas out of 13 distinct ideas, which was acceptable for deployment (cf. Schulz et al., 2018, 2019). While the model performed well for frequent idea types, it is challenging to achieve high accuracy across all 13 ideas from our relatively small dataset. Five ideas were either unobserved or appeared less than 5 times in our dataset: 1) Energy is created by plants; 2)Plants transform light energy into glucose; 3) Plants store energy in glucose; 4) Plant releases energy from glucose to grow; 5) Animal uses the glucose or food to grow or move.

Given these limitations, I restricted the deployed version of the model to the eight most reliably detected ideas with their RPP-designed guidance. The AI adaptive dialog was deployed in the start-of-year inventory. There were two rounds of adaptive guidance in the dialog. Both rounds aimed to elicit further reasoning, with prompts such as "What else…?" or "What happens next…?" If students repeated the same idea in the second round, the system provided a distinguishing prompt, such as "How is this different from…?" or "What is the process…?"

If multiple ideas are detected, the dialog responds to intuitive ideas first to address misconceptions. Among normative ideas, a typical response explains energy and matter transfer from the sun to the plants, and then from plants to animals. Therefore, if multiple normative ideas are present, the dialog first responds to ideas about animal cellular respiration (assuming they have already explained ideas about photosynthesis and energy transfer), then ideas about energy transfer from plants to animals, and then ideas about photosynthesis. This ordering assumes that students who mention later-stage ideas (e.g., animal cellular respiration) have likely already addressed earlier-stage concepts (e.g., energy transfer from plants to animals). For example, if a student expressed an idea related to photosynthesis, the prompt would aim to extend their explanation by eliciting how that energy reaches animals, encouraging them to consider concepts related to cellular respiration.

This first iteration was implemented during the start-of-year assessment in Fall 2022.

3.4.2 Iteration 2 - Researchers added distinguishing guidance following KI to help students elaborate ideas in the dialog

Following the initial classroom inventory implementation, I collected and reviewed student dialog logs with the RPP from Iteration 1, including students' initial and revised explanations, NLP-detected ideas, and Knowledge Integration (KI) scores. A more detailed quantitative analysis is reported in the Results section. To inform design revisions, a sample of 20 student dialogs was randomly selected and reviewed in two small groups consisting of five researchers, three science teachers, two software designers, and one computer scientist. I analyzed these dialogs through the KI framework: were students being supported to distinguish among ideas? Were the prompts helping them connect concepts across the energy flow pathway?

Across both groups, we found that second-round eliciting guidance often led to minimal elaboration. Students frequently responded with "I don't know" or repeated earlier ideas. However, distinguishing guidance in the second round that asked students to contrast or clarify ideas facilitated elaboration. For example, as shown in Figure 5, Student M received an eliciting prompt in Round 2: *"What type of energy do animals get from plants?"* The student replied, *"I just told you they get food."* This does not add any new idea to the dialog. In contrast, Student A received a distinguishing prompt: *"How do plants get energy from the sun differently than animals?"* This question prompted Student A to articulate the idea of plant photosynthesis more clearly.

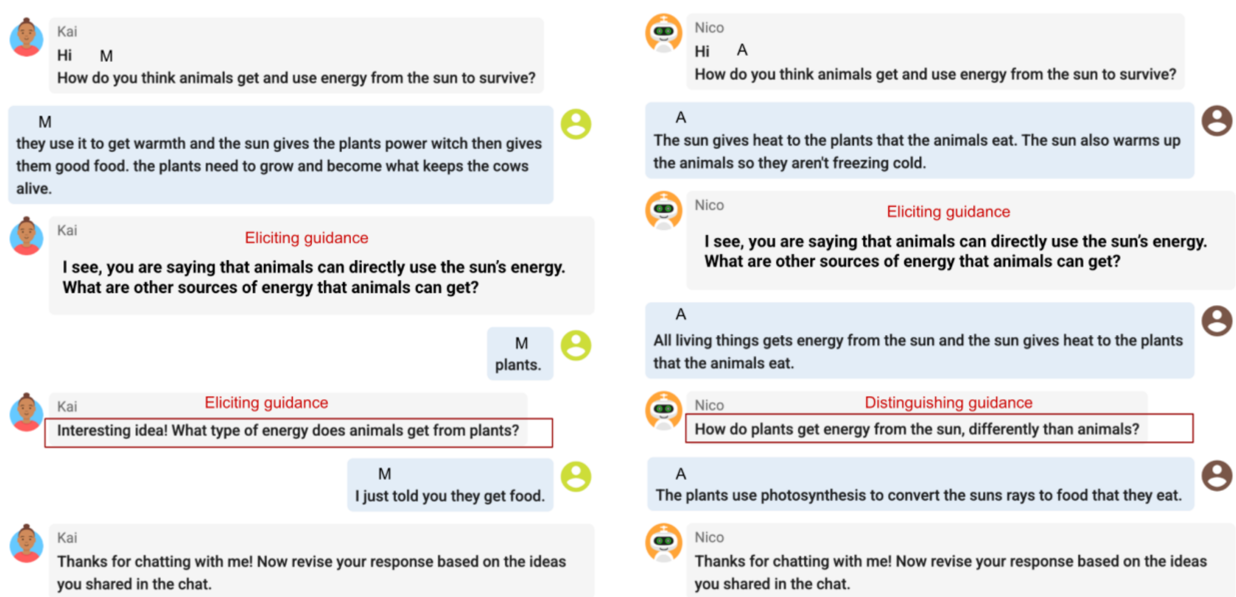

Figure 5. Examples of student dialogs at Iteration 1. Student M receives an eliciting prompt in the second round (on the left) and Student A receives a distinguishing prompt in the second round (Student A on the right).

These findings align with prior research showing that prompts encouraging students to compare ideas or evaluate alternatives are more effective for deepening understanding than those that simply elicit elaboration. For example, Matuk and Linn (2018) used a digital "Idea Manager" to ask students to either choose a peer's idea that differed from their own or one that reinforced their own idea. They found that students who were guided to confront ideas that contradicted their own thinking showed greater knowledge integration gains than those who only saw ideas similar to theirs. Clark and Sampson (2007) found out that guidance on comparing competing ideas improved the quality of arguments and understanding in online discussion. Davis (2003) found that students given only generic “think about it” prompts often reflected superficially, while guidance in how to reflect (e.g. prompting specific distinctions or connections) guided students to integrate ideas more productively. Other researchers also found that elaborating prompts guided students to explain only the accepted scientific theory while the competing prompts had students differentiate between two theories (Donnelly et al., 2016).

In response, I revised the dialog guidance in Iteration 2 to exclusively use distinguishing prompts in the second round. This updated version was implemented at the end-of-year assessments in Spring 2023.

3.4.3 Iteration 3 - Teachers reviewed and expanded the rubric to 19 ideas

After the classroom runs at Iteration 2, the lead researcher randomly selected 20 student dialogs that represent a range of KI levels and frequent ideas. These were reviewed by two small groups consisting of 4 researchers, 1 pre-service science teacher researcher, 2 software designers, and 1 computer scientist. In both groups, they identified ideas that were not detected by the existing model or guidance that were assigned incorrectly. For example, one response, “The cow eats plants to get their food then digest it”, was incorrectly labeled as energy transfer, highlighting the difficulty in distinguishing between students describing food as matter versus food as a source of energy. Another response, “energy decreased down the cascade of food chain”, has no ideas detected but they mentioned an important rule of energy transfer. Therefore, the RPP team decided to further validate the rubric with partner teachers.

Iterating on the Rubric. To refine the original 13-idea rubric, three science teachers in our RPP, Mary, Dan, and Jack, who all had over 10 years of experience teaching WISE units including the Energy Story Activity, were recruited through our RPP summer workshop. Using another set of 20 dialogs, teachers identified six additional ideas commonly expressed by students but absent from the rubric (Figure 6). These included entry-point ideas for building understanding of photosynthesis (e.g. “sun helps plants survive” and “sun helps animals survive”). They also noticed common ideas from their prior instruction (e.g., the “10% rule in a food chain energy pyramid”, and “the total energy in the food chain does not change”). They also identified two intuitive ideas (e.g., “in the food chain. Fox eats rabbits who eat grass, then fox gets the original energy”, and “too much sun will cause skin cancer”). These insights led to the expansion of the rubric from 13 to 19 ideas (see Figure 6). The expanded rubric better supported the full range of KI processes: new entry-point ideas (e.g., "sun helps plants survive") gave students more accessible starting points for eliciting guidance, while curriculum-specific ideas (e.g., "10% rule in food chain") allowed the dialog to connect to what students had learned in instruction.

Original idea	Refined idea description
1-Reactants of Photosynthesis	1a-Reactants of Photosynthesis such as CO2 and H2O (refined)
2-Products of Photosynthesis	2a-Products of Photosynthesis such as glucose and O2(refined)
3-Photosynthesis Chemical Reaction	3a-Photosynthesis Chemical Reaction (refined)
4-Energy Transformation in photosynthesis	4a-Energy Transformation in photosynthesis (refined)
5-Energy is Created	5a-Energy is Created
6-Energy becomes Matter	6a-Energy becomes Matter
7-Plant Cellular Respiration	7a-Plant use the energy they get to grow
8-Plant Stores Energy	8a-Plant Stores Energy (refined)
9-Energy Transfer from plants to animals	9a-Energy Transfer from plants to animals (refined)
10-Animal Cellular Respiration	10a-Animal Cellular Respiration (refined)
11-Animal use glucose to grow/run/hunt	11a-Animal use the energy they get to grow/run/hunt
12-Animals directly use Sun's energy	12a-Animals Use the Sun for warmth/vitamins (refined)
13-Animal Eat Plants for Food/Nutrition	13a-Animal Eat Plants for Food/Nutrition (refined)
	14-Sun Helps Plants grow
	15-Sun helps Animals grow
	16-Energy Conservation
	17-Energy decrease down the cascade of food chain
	18-Animals use glucose without any energy transformation
	19-Negative impact of the Sun on Animals

Figure 6. The original idea rubric (13 ideas, left) and refined idea rubric (19 ideas, right). Ideas in *Italic were added by teachers.*

Refining the Idea Rubric. After expert teachers expanded the rubric, two researchers (the lead researcher who designed the original rubric and one pre-service science teacher researcher) iteratively revised the expanded rubric through four rounds of annotation using 20% of student dialogs from Iteration 2. In the first round, we added distinguishing criteria to clarify overlapping ideas. For instance, we added rules to differentiate the idea of animals eating plants for energy (idea 9) from the idea of animals eating plants for food and nutrition (idea 13): Idea 13 was used when students mentioned non-fuel matter such as fibers or nutrients related to digestion, while Idea 9 was used when students referred to energy transfer, including terms like glucose, carbohydrates, or energy itself. In the second round, we reorganized the 19 ideas into three categories: normative (scientifically accurate), intuitive (based on personal experience), and broad (general or vague entry points), enabling the dialog to scaffold a wider range of student responses (Appendix Table 2). The third round focused on narrowing definitions for low-frequency ideas and recategorizing ambiguous cases, such as redefining Idea 7 (plant cellular respiration). Many students mentioned plant growth without explicitly referencing glucose or cellular respiration, so we refined the idea as "plant use the energy they get to grow" to support more accessible entry points. In the fourth round, two researchers reached high inter-rater reliability (Cohen's Kappa = 0.81).

After validating the updated rubric, the researchers annotated the remaining 80% of dialogs from Iteration 2. The newly labeled dataset was used to retrain the NLP model, which achieved a micro-averaged F1-score of 0.7297, which was acceptable for deployment (cf. Schulz et al., 2018, 2019). Adaptive guidance was also added for the reliably detected new ideas (F1-score > 0.67). The guidance was then reviewed by the three expert teachers separately, and revisions were made to emulate their scaffolding to students or asking for elaboration. The rubric and guidance were finalized in a RPP team meeting.

The Iteration 3 dialog was implemented in classrooms during the start-of-year inventory in Fall 2023, and the end-of-year inventory in Spring 2024.

3.4.4 Iteration 4 - Researchers revised the dialog to guide on most promising idea first and used distinguishing guidance in both rounds

In earlier iterations, the guidance prioritized responding to intuitive ideas when multiple ideas were detected, based on the assumption that addressing misconceptions first could support conceptual growth. However, our reviews during Iteration 3 revealed limitations to this approach.

I collected and analyzed dialog data from Iteration 3, including initial and revised student explanations, KI scores, detected ideas, and adaptive guidance. A random sample of 20 student dialogs was reviewed in two small groups composed of five researchers, three science teachers, two software designers, and one computer scientist.

During these reviews, we observed that first-round eliciting prompts often did not support productive elaboration, especially when students had already articulated multiple ideas. In some cases, prompts directed students back to their intuitive ideas rather than building on the more promising normative ones. For example, as shown in Figure 7, student AT expressed two ideas in their initial response: a normative idea about energy transfer from the sun to plants and animals, and an intuitive idea that animals directly use the sun's energy to get vitamins. The thought buddy responded to the intuitive idea with an eliciting prompt: "What are other sources of energy that animals can get from their food?" Because AT had already mentioned the mechanism, they responded with a vague new idea: "the energy of heat transfer and things from the air." This response did not deepen their explanation of the energy transfer process.

Moreover, both review groups noted that distinguishing prompts, those that asked students why or to compare or clarify ideas, consistently facilitated more meaningful elaboration, even in the second round when students were running out of ideas. In AT's second-round response, the distinguishing prompt helped them return to and extend their scientifically accurate explanation.

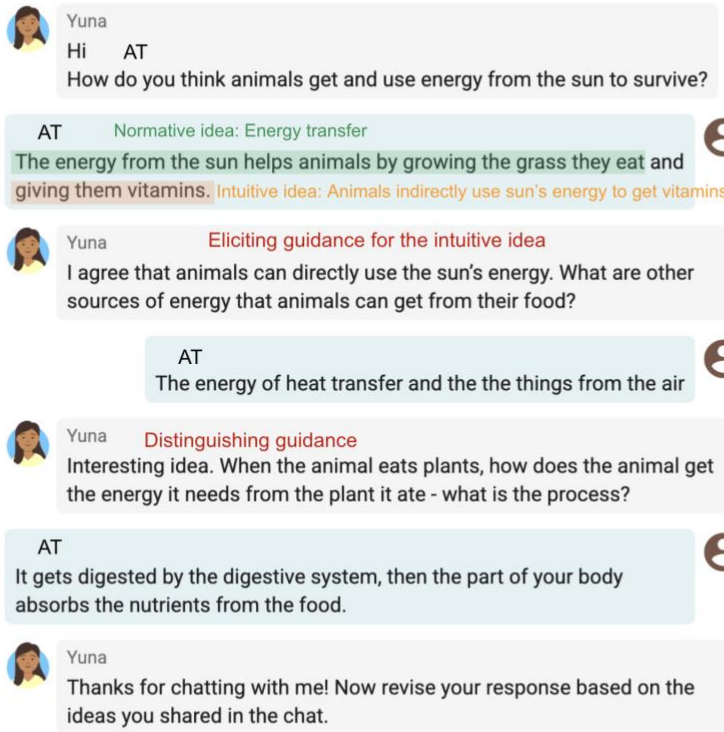

Figure 7. Examples of student dialogs at Iteration 3. They received an eliciting prompt based on their intuitive idea in the first round and a distinguishing prompt in the second round.

These findings align with prior research and learning theory. Rather than focusing on the deficit framing that evaluates what students do not know and cannot do based on diagnosing and fixing error (Valencia, 1997), effective guidance often builds on what students already know: making their ideas visible (Linn & Hsi, 2000), valuing their diverse perspectives (Bang & Medin, 2010), and intentionally connecting them to scientific phenomenon (Rosebery et al., 2016). This also aligns with KI pedagogical framework for connecting ideas to evidence and for linking one idea to another, rather than punishing having intuitive ideas or rewarding the accumulation of isolated ideas (Linn & Eylon, 2011).

To address this, I revised both our guidance type and prompting strategy in ways that better aligned with KI principles of building on students' existing knowledge. First, I shifted to using distinguishing guidance in BOTH rounds (not just Round 2), recognizing that prompts asking students to compare, explain mechanisms, or clarify relationships consistently supported deeper integration than eliciting prompts that simply asked for more information. This change reflects KI's emphasis on helping students distinguish among ideas and connect them to evidence (Linn et al., 2023), rather than merely accumulating additional ideas.

Second, I revised our prompting strategy to prioritize the most conceptually sophisticated normative idea when students expressed multiple ideas. Rather than defaulting to addressing intuitive ideas first (a deficit-oriented approach that focuses on what students don't yet understand), this revision built on what students already knew by extending their most promising scientific thinking. This aligns with KI's principle of treating all student ideas as resources: when students have already articulated a mechanistic understanding of energy transfer, the dialog should help them deepen and connect that understanding rather than redirecting them to less developed ideas.

This change was particularly important for students who began with multiple ideas, as it allowed the dialog to support progressive integration rather than fragmentation of their thinking.

These revisions to guidance and prompting strategy were developed collaboratively with two science teachers (Amy and Ben) who helped refine the distinguishing prompts for clarity and classroom appropriateness. The updated dialog was implemented in Fall 2024 (Iteration 4).

4 Methods

4.1 Participants

Across four design iterations, the AI adaptive dialog was implemented as part of a benchmark assessment (Inventory) run in classrooms without prior instruction. Table 2 summarizes the key features of each iteration, including the NLP model, guidance and prompting strategies, and the number of students who participated.

Table 2. Dialog iterations and participants

Iterations	NLP Model	Guidance Design	Prompting Design	Participants
Iteration 1	Original (13 ideas)	Round 1: elicit Round 2: elicit unless idea repeated	Guide on intuitive ideas first	896 students in start-of-year inventory Fall 2022
Iteration 2	Original (13 ideas)	Round 1: elicit Round 2: distinguish	Guide on intuitive ideas first	745 students in end-of-year inventory Spring 2023
Iteration 3	Expanded (19 ideas)	Round 1: elicit Round 2: distinguish	Guide on intuitive ideas first	691 students in start-of-year inventory Fall 2023; 420 students in end-of-year inventory Spring 2024
Iteration 4	Expanded (19 ideas)	Round 1: distinguish Round 2: distinguish	Guide on normative ideas first	647 students in start-of-year inventory Fall 2024

In total, 3,399 students participated across all four iterations. Sample sizes per iteration ranged from 420 to 896 students, with data drawn from multiple middle schools across the western United States.

Across the four iterations, 15 teachers implemented the Inventory in their classrooms. Five of these teachers were members of our RPP and had been deeply involved in the co-design process (Table 1, Section 3.2). The remaining 10 teachers were recruited specifically to run the Inventory in their schools, with some participating in a single iteration and others across multiple years. Teachers outside the RPP were offered compensation for running the Inventory. Many teachers were also motivated by the assessment value: the Inventory provided insights into their students' baseline understanding (KI levels, frequent ideas) at the start of the year and allowed them to track development by end of instruction. This combination of support and pedagogical relevance enabled us to implement the dialog across diverse school contexts in the western United States.

4.2 Data Analysis

4.2.1 Data Segmentation for Analysis

To examine the impact of iterative design changes on student learning, I conducted two parallel strands of analysis: 1) changes in students' Knowledge Integration (KI) scores, and 2) changes in the number and type of ideas students added during the adaptive dialog. These outcomes reflect different dimensions of learning, therefore required different data segmentation strategies and statistical models.

For analytical purposes, I defined two units of analysis. A 'dialog' refers to the complete interaction: one student's initial explanation, two adaptive prompts with responses, and their revised explanation. A 'round' refers to one prompt-response cycle within the dialog. KI score analyses used the dialog as the unit of analysis (comparing initial vs. revised explanations), while idea change analyses used rounds as the unit of analysis to examine how students responded to each prompt.

For each design feature (guidance type, NLP idea rubric, and prompt strategy), I isolated relevant iterations while holding other factors constant (e.g., timing of implementation, NLP version). Table 3 summarizes the analyses of KI score gains across design iterations.

Table 3. Dataset Segments Used for Knowledge Integration (KI) Score Analyses

Section	Design Feature Analyzed	Iterations Used	Sample Size (N)	Outcome Variable	Model Used
5.1	Overall Dialog Impact	1, 2, 3, 4	3399 student dialogs	KI scores (initial vs revised)	CLMM
5.2	Guidance Type (Elicit vs Distinguish)	1, 2	1641 student dialogs	KI scores (initial vs revised)	CLMM
5.3	NLP Idea Rubric (13 vs. 19 ideas)	2 (Spring), 3 (Spring)	1165 student dialogs	KI scores (initial vs revised)	CLMM
5.4	Prompt Strategy (Intuitive vs. Normative First)	3 (Fall), 4 (Fall)	1338 student dialogs	KI scores (initial vs revised)	CLMM

Table 4 outlines analyses of idea change patterns in student responses across rounds. In analyses of idea development, I further segmented rounds based on students' initial idea expression (e.g., whether their prior response included zero, one, or multiple ideas), as this shaped how they interacted with dialog guidance.

Table 4. Dataset Segments Used for Idea Addition Analyses

Section	Design Feature Analyzed	Iterations Used	Subset Criteria	Sample Size (N)	Outcome Variable	Model Used

5.2	Guidance Type (Elicit vs Distinguish)	1, 2	Rounds with no initial idea at start of round	929 rounds	New ideas, New normative ideas	GEE
			Rounds with idea(s) at the start of round	2353 rounds		
5.3	NLP Idea Rubric (13 vs. 19 ideas)	2 (Spring), 3 (Spring)	All student dialogs	1165 dialogs	Key idea presence (binary)	GEE
5.4	Prompt Strategy (Guide on Intuitive vs. Normative First)	3 (Fall), 4 (Fall)	Round 2 where students start with two or more ideas	360 rounds	New ideas, New normative ideas	GEE

4.2.2 Statistical Modeling to Evaluate Iteration Effects

To systematically evaluate whether design changes improved student learning outcomes, I conducted quantitative analyses using the full dataset from all implementations. These analyses served a different purpose than the RPP collaborative review (Section 3.3.6): rather than identifying specific design problems, these statistical tests evaluated whether observed differences in student outcomes across iterations were reliable and not due to chance variation or confounding factors.

Analyzing Knowledge Integration Score Changes

To analyze changes in students' KI (Knowledge Integration) scores across design iterations in the dialog, I used Cumulative Link Mixed Models (CLMMs). KI scores were rated on a 5-point scale, where higher scores indicated more integrated understanding of energy transfer and transformation in the food chain. For example, For example, the difference between a KI score 1 (off-topic) and KI score 2 (intuitive or inaccurate ideas) may not represent the same magnitude of change as between KI score 3 (partial link with one normative idea) and KI 4 (one scientifically complete and valid connection between normative ideas). Because these scores reflect ordered categories without assuming equal intervals between levels, CLMMs are more appropriate than linear models. This approach models the cumulative probability of achieving a certain KI level or higher, aligning with the rubric's qualitatively different and ordered levels. CLMMs in this study also included random intercepts for students and schools to account for repeated measures nested within students, and students nested within schools. The models have fixed effects for dialog round (initial vs revised explanations), iterations, and demographic covariates (language background, grade, and gender). This allowed us to examine how students' revised their explanations across dialog and design iterations, while accounting for demographic and contextual differences.

Analyzing Idea Incorporation Patterns

Generalized estimating equation (GEE) models were applied to analyze the effects of design iterations and dialog on student idea changes. I used GEE models because the dependent variable, number of new ideas, is count data and not normally distributed. GEE can handle repeated measures of non-normal data and provides robust estimates of average effects at the population level. Given the skewed distribution and high frequency of zeros for the new ideas added, GEE provides robust standard errors and interpretable estimates of how different design revisions influenced idea development on average.

Controlling for Confounding Variables

All models included covariates to control for potential confounding: Language background (Multilingual students may face additional linguistic demands that affect both idea expression and comprehension of guidance), grade level (Students in 6th, 7th, and 8th grade have different amounts of prior science instruction), school (schools varied in demographics (e.g., 86-95% free/reduced lunch across schools)) and unmeasured factors. Including these covariates allowed me to isolate the effects of design changes while accounting for known sources of variation in student outcomes.

It is important to note that these statistical analyses served a different purpose than the collaborative RPP review process described in Section 3.3.6. The RPP reviewed purposive samples of 20 dialogs per iteration to identify specific design problems and generate solutions grounded in teacher expertise. In contrast, the quantitative analyses reported in Section 5 used the complete dataset (N = 3,399 student dialogs across all four iterations) to evaluate whether those design solutions produced statistically reliable improvements in student learning outcomes. These quantitative analyses were conducted by me (the lead researcher) and reviewed with the core education research team.

5 Results

This section begins with an overview of students' initial and revised KI scores before and after the dialog across four iterations. This provides important context for interpreting the outcomes of each iterative change. Following this overview, the results are organized into three key comparisons, each aligned with a specific design decision made during the iterative development of the AI dialog. These comparisons focus on how changes in 1) guidance design, 2) NLP idea rubric, and 3) prompting strategy influenced students' KI scores and idea incorporation.

5.1 Overview of students' KI scores across all four iterations

As shown in Figure 8, in each of the four iterations, students' KI scores improved from initial to revised explanations. The average gain (Δ) was 0.22 in Iteration 1 (Fall 2022, $N = 896$, $t(895) = 9.13$, $p < .001$, $d = 0.30$), KI gain was 0.33 in Iteration 2 (Spring 2023, $N = 745$, $t(744) = 11.02$, $p < .001$, $d = 0.40$), KI gain was 0.35 in Iteration 3 in Fall 2023 ($N = 691$, $t(690) = 10.85$, $p < .001$, $d = 0.41$), 0.49 in Iteration 3 in Spring 2024 ($N = 420$, $t(419) = 9.17$, $p < .001$, $d = 0.45$), and 0.35 in Iteration 4 (Fall 2024, $N = 647$, $t(646) = 10.52$, $p < .001$, $d = 0.41$). Fall inventory was administered at the start of the school year, while Spring inventory took place at the end of the year, which may have influenced baseline student scores. Moreover, I have three partner teachers who taught the photosynthesis WISE unit in December 2023 to April 2024. The Energy Story assessment item is embedded in the curriculum unit. This resulted the higher initial KI scores and more improvements in Spring 2024 inventory. While initial KI scores were generally higher in the spring inventories, the increased gain observed from Iteration 1 to Iteration 3, and maintained in Iteration 4 despite it being a fall administration, suggests that the iterative design changes may have contributed to students' improved ability to revise their explanations.

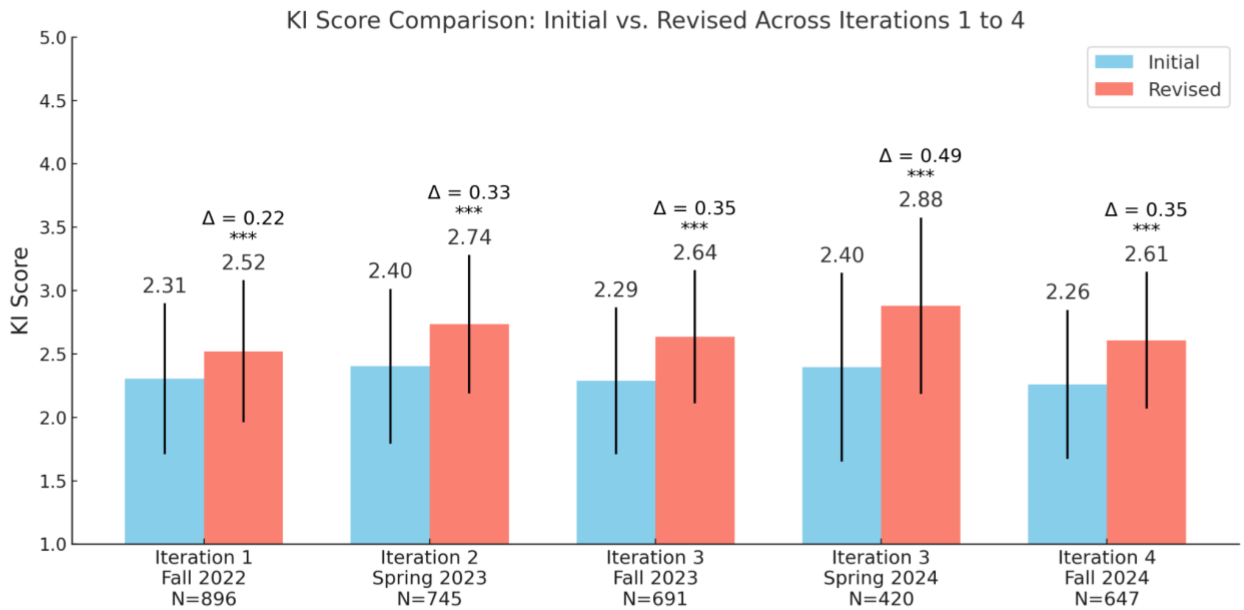

Figure 8. Mean Knowledge Integration scores at initial and revised explanations across Iterations 1–4. Note. Error bars represent ±1 SD. Sample sizes shown below each iteration. Mean gains (Δ) shown above each pair. ***p < .001 (paired t-tests).

Because each iteration introduced different design features, including changes to guidance type, idea detection rubric, and prompting strategy, direct comparison of effect sizes across iterations would conflate design improvements with contextual differences. Therefore, the following sections present controlled comparisons isolating specific design features: Section 5.2 examines guidance type by comparing Iterations 1 and 2 (which used the same NLP model and prompting strategy), Section 5.3 examines rubric expansion by comparing Iterations 2 and 3 Spring (both end-of-year administrations with the same guidance structure), and Section 5.4 examines prompting strategy by comparing Iterations 3 and 4 Fall (both start-of-year administrations with the same expanded model).

5.2 Guidance Type: Comparing Eliciting and Distinguishing Guidance

To examine how different guidance strategies influence students' incorporation of normative ideas and improvements in knowledge integration (KI) scores, I compared data from Iteration 1 and Iteration 2. These iterations offered a clear contrast in guidance types while maintaining the same NLP idea detection model (the original model with 13 ideas) and prompting strategy (guiding intuitive ideas first). Specifically, Iteration 1 employed primarily eliciting guidance, providing distinguishing guidance only when students repeated ideas in the second round. In contrast, Iteration 2 consistently provided distinguishing guidance during the second round. This comparison allowed me to isolate the effects of guidance type (eliciting versus distinguishing) and assess their distinct impacts on student learning outcomes.

5.2.1 Knowledge Integration (KI) Score Changes by Guidance Type

To examine the impact of guidance type on students' knowledge integration (KI), I compared KI scores at initial and revised explanations in Iteration 1 (Round 1: Eliciting guidance; Round 2: Eliciting guidance, with Distinguish guidance provided if the same idea was repeated from Round 1) with Iteration 2 (Round 1: Eliciting guidance; Round 2: Distinguishing gudiance) using a

Cumulative Link Mixed Model (CLMM). The model included fixed effects for Dialog Round (Initial vs. Revised), Iteration (1 vs. 2), Language background (Monolingual vs. Multilingual), and Grade (6, 7, 8), with random intercepts for Student ID and School to account for clustering. The model fit was acceptable (AIC = 5067.52, n = 3282). Likelihood ratio tests confirmed the importance of several predictors. Removing the Dialog Round and Iteration interaction significantly reduced model fit (ΔAIC = 17, LRT = 19.4, $p < .001$). Similarly, removing Grade level reduced model fit (ΔAIC = 42, LRT = 46.2, $p < .001$), suggesting that although grade effects were not significant, grade accounted for meaningful variation in the model.

Results showed that in both iterations, adaptive dialog significantly improved students' KI scores from their initial to revised explanations ($\beta = 1.13$, $SE = 0.116$, $z = 9.76$, $p < 0.001$). Students also received higher KI scores in Iteration 2 dialog compared to Iteration 1 ($\beta = 0.706$, $SE = 0.136$, $z = 5.21$, $p < 0.001$). This difference could reflect the timing of implementation: Iteration 2 occurred during an end-of-year inventory, whereas Iteration 1 was implemented at the start of the academic year. Thus, some students in Iteration 2 might have learned relevant science concepts, such as photosynthesis and cellular respiration during the year, influencing their initial scores. Grade and language background effects were not statistically significant. Random effect variances indicated substantial individual-level variability (σ^2_ID = 3.29) and moderate school-level variability (σ^2_School = 0.82).

Importantly, a significant interaction between dialog and iteration indicated that the improvement in KI scores from the initial to the revised explanations was greater in Iteration 2 ($\beta = 0.748$, $SE = 0.171$, $z = 4.38$, $p < 0.001$). Students in Iteration 2 were 2.1 times more likely to improve their KI scores from the initial to the revised response compared to students in Iteration 1. Overall, these findings suggest that incorporating distinguishing guidance at Round 2 (Iteration 2) more effectively supported students in integrating scientific ideas than the primarily eliciting approach used in Iteration 1. The statistical model controlled for potential confounding variables, including student linguistic background, grade, gender, and school, and accounted for repeated measures through a random intercept for each student.

5.2.2 Changes in Student Ideas by Guidance Type

To examine the impact of guidance type on students' ideas, I compared student total new ideas and new normative ideas at Iteration 1 (Round 1: Eliciting guidance; Round 2: Eliciting guidance, with Distinguish guidance provided if the same idea was repeated from Round 1) with Iteration 2 (Round 1: Eliciting guidance; Round 2: Distinguishing guidance) using generalized estimating equation (GEE) models with Poisson distribution. New idea means that the student added a new idea in their response to the prompt compared to their previous response, new normative idea means the student added a new normative idea in their response to the prompt. To better understand the impact of our guidance type, I divided the data into two sets: 1) students who did not have any detectable ideas in the initial response; 2) students who had at least one idea expressed in their initial response.

When students don't have any detectable ideas shared in their response, or their responses were identified as "nonscorable" by the model, they received two generic prompts ($N_{round} = 929$). The eliciting prompt asks, "Can you tell me more about this idea or another one in your explanation? I am still learning about student ideas to become a better thought partner.", the distinguishing prompt asks, "How does the energy transfer from the sun to the plants and then to the animals?". A generalized estimating equation (GEE) model with a Poisson distribution was used to examine the effect of guidance type and other variables on the number of new ideas added by students. The overall model included guidance type (elicit vs distinguish) as the fixed effect, dialog round, language background, grade, gender, and school were included as covariates. Results indicated a marginally significant effect of guidance type while controlling for other variables, with students in the elicit condition adding more new ideas than those in the distinguish condition ($\beta = 0.23$, *SE* = 0.13, *Wald* = 3.07, $p = 0.078$, Figure 9 left). No significant effects were observed for round, language, grade, gender, or school. I then used another Poisson Generalized Linear Model to examine the effect of guidance type and other variables on the number of new normative ideas added by students. Results indicated no significant effect of guidance type, round, language, grade, gender, or school. I included the Mean and SD in the Figure 9 (right).

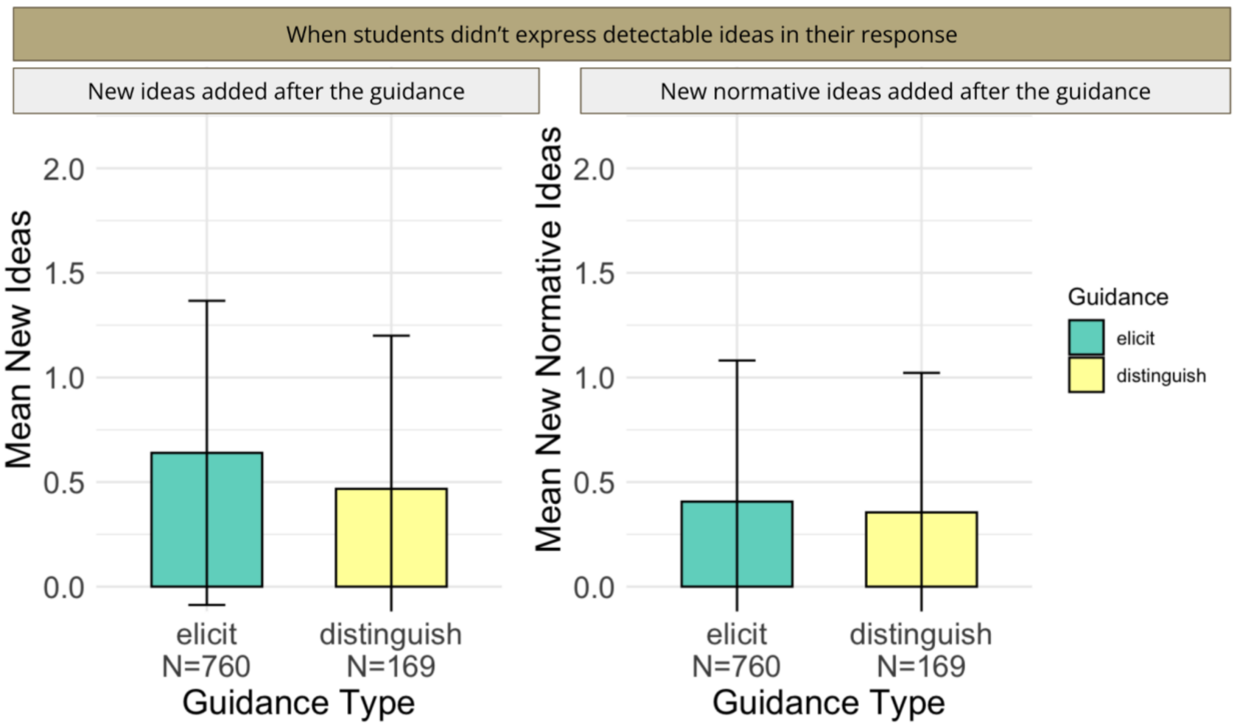

Figure 9. Number of new ideas and new normative ideas added by students with no initial detectable ideas, by guidance type (elicit vs. distinguish)

When at least one idea was detected in their responses, two rounds of adaptive prompts were assigned to students ($N_{round} = 2353$). Two generalized estimating equation (GEE) models were conducted to examine how guidance type influenced the number of new ideas and new normative ideas students added in their responses when they shared at least one detectable idea. Both models included dialog round, language, grade, gender, and school as covariates.

In the first model predicting the total number of new ideas, students in the elicit condition added significantly fewer ideas than those in the distinguish condition ($\beta = -0.47$, SE = 0.20, *Wald* = 5.42, $p = 0.020$). In the second model predicting the number of new normative ideas, students in the elicit condition again added significantly fewer normative ideas compared to those in the distinguish condition ($\beta = -0.55$, SE = 0.25, *Wald* = 4.74, $p = 0.029$). Across both models, no significant effects were found for language background, grade level, school, or gender.

The results indicated that after students shared their initial relevant idea, students who received distinguishing prompts added significantly more new ideas and more new normative ideas (Figure 10). Due to the nature of dialog, students added fewer ideas in their second round.

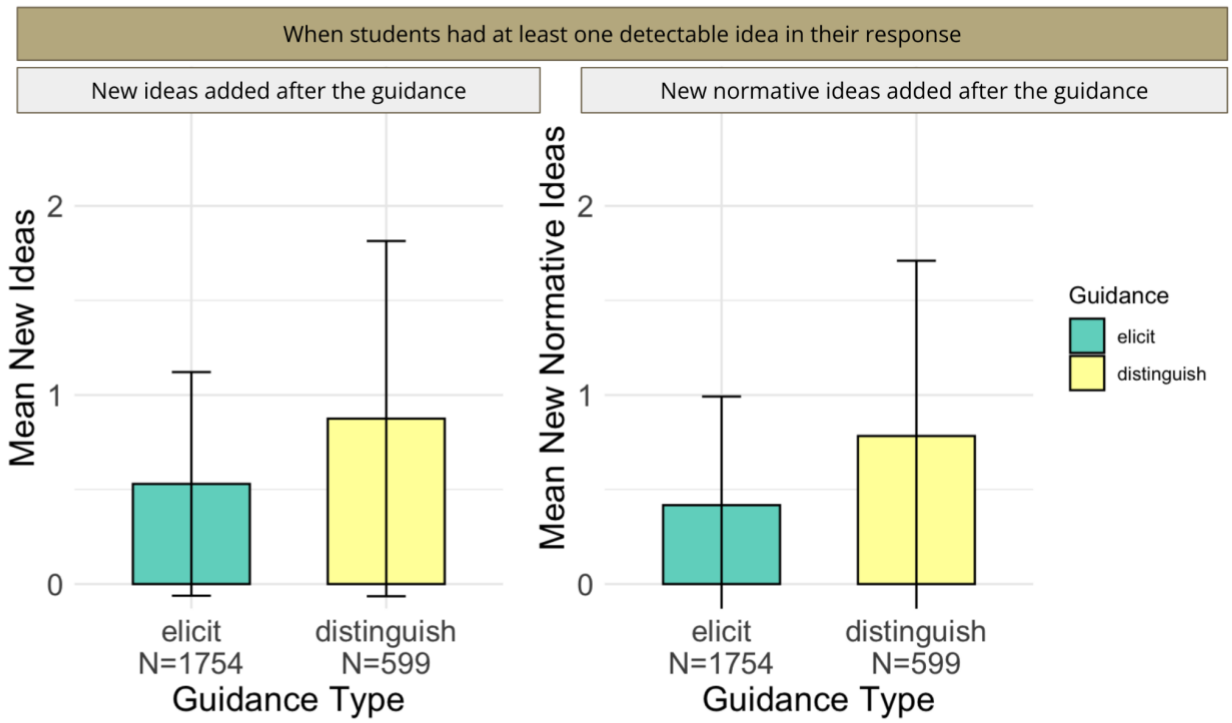

Figure 10. Number of new ideas and new normative ideas added by students with initial detectable ideas, by guidance type (eliciting vs. distinguishing)

In summary, the results comparing Iteration 1 and Iteration 2 showed that eliciting prompt helped students add more ideas when they don't have anything to share at the beginning. The distinguishing prompt helped students add more new ideas and normative ideas when they shared initial knowledge at the beginning. For example, Table 3 illustrates how students responded to adaptive prompts targeting the idea of animals directly using the sun's energy for warmth or vitamins. In the first example, a student initially stated that animals rely on the sun to avoid freezing. When prompted with an elicit question, the student expanded their reasoning by linking sunlight to animals' food sources (intuitive idea of animals eating plants for food and nutrition). According to the data summary, the most added idea after this eliciting prompt ($N_{total} = 271$, $N_{idea12added} = 134$) is an intuitive idea of animals eating plants for food and nutrition.

In the second example, the student initially claimed that the sun provides warmth and vitamins. After receiving a distinguishing prompt, the student shifted toward a more accurate link to energy, stating that animals obtain energy indirectly by eating food that collects energy from the sun. According to our data summary, the most added idea after this distinguishing prompt is a normative idea of photosynthesis ($N_{total} = 58$, $N_{idea12added} = 30$).

Table 3. Student dialog examples of how they responded to the adaptive prompts for idea 12

For idea 12 (Animals directly use Sun's energy for warmth or/and vitamins)			
Example	Student initial response	Prompt	Student response
1	they get heat from the sun to not freeze to death	Elicit: I agree that animals can directly use the sun's energy.	the sun also keeps their food alive, like

		What are other sources of energy that animals can get from their food?	animals, insects and greens.
2	because the sun keeps them warm and gives them vitamins	Distinguish: The sun helps animals in lots of ways! How do you think the sun helps animals get energy from food?	I think the food collects the energy from the sun by photosynthesis and then the animals eat the food

5.3 Rubric Expansion: Comparing Expanded and Original NLP Idea Detection Model

To examine the impact of rubric expansion on students' knowledge integration (KI), I compared KI scores at initial and revised explanations in Iteration 2 (Initial idea rubric with 13 ideas) with Iteration 3 (Expanded idea rubric with 19 ideas) using CLMM with a logit link using the Laplace approximation. To control for the effect of semesters, I picked the Spring 2024 inventory data (N = 420) in Iteration 3 to compare with the Spring 2023 inventory data (N = 745) in Iteration 2. The model included fixed effects for Dialog Round (Initial vs. Revised), Iteration (2 vs. 3), Language background (Monolingual vs. Multilingual), and Grade level (6, 7, 8), with random intercepts for Student ID and School to account for nesting and individual variability. The model fitted well (AIC = 3964.75, n = 2330). Substantial variability was observed at the individual level ($\sigma^2 = 5.21$) and moderate variability at the school level ($\sigma^2 = 0.83$), supporting the inclusion of both random effects. Likelihood ratio tests confirmed the importance of several predictors. Removing the Dialog Round and Iteration interaction significantly reduced model fit (ΔAIC = 11, LRT = 12.50, $p < 0.001$). Similarly, removing Language effect reduced model fit (ΔAIC = 3, LRT = 4.85, $p = 0.028$), while removing Grade factor had no significant impact.

The analysis revealed a significant main effect of the dialog: students earned higher KI scores on the revised responses compared to initial responses ($\beta = 1.86$, SE = 0.145, $z = 12.82$, $p < 0.001$). There was also a significant interaction between dialog and iteration ($\beta = 0.75$, SE = 0.212, $z = 3.51$, $p < 0.001$), indicating that the effect of dialog varied across design iterations. Students in Iteration 3 were 2.1 times more likely to improve their KI scores after the dialog. The expanded idea detection model more effectively supported knowledge integration than the initial idea model. I also found that multilingual students had 33% lower odds of receiving a higher KI score compared to monolingual students ($\beta = -0.40$, SE = 0.18, $z = -2.20$, $p = 0.028$).

Due to the expanded idea rubric in Iteration 3, I did not conduct statistical comparisons of total idea change between Iterations 2 and 3, as the outcome variable was not measured equivalently. Instead, I focused on two key ideas critical to understanding energy transfer and transformation in ecosystems: energy transfer and animal cellular respiration. Generalized Estimating Equation (GEE) logistic models revealed a significant interaction between dialog (initial vs revised explanation) and Iteration ($\beta = 0.75$, $p < 0.001$), indicating that students in Iteration 3 are more likely to add the energy transfer idea after the dialog than students in Iteration 2. A similar pattern emerged for the idea of animal cellular respiration, with a significant interaction between dialog (initial vs revised explanation) and Iteration for the idea of animal cellular respiration ($\beta = 1.565$, $p < 0.001$). Together, these findings suggest that the expanded idea rubric in Iteration 3 may have supported students in incorporating more mechanistic explanations into their revised responses.

The findings underscore that expanding the NLP idea detection model significantly enhanced students' knowledge integration during the adaptive dialog, highlighting the importance of incorporating teacher input into the adaptive dialog design. For example, the following qualitative analysis of student responses demonstrates how different dialog models supported varying levels of elaboration and scientific reasoning. Lucia, using the initial model (Figure 11, left), began with an energy transfer idea and, after receiving adaptive guidance, expanded her original explanation with an observation "animals eat plants and then poop it out." While this elaboration provided more detail, it remained at an everyday language level, and Lucia did not incorporate scientific mechanisms like cellular respiration. In contrast, Skye, using the expanded model (Figure 11, right), started with a clear description of energy transfer and the word "photosynthesis". The expanded model's specific guidance helped her add more scientific depth, introducing cellular respiration and using terms like "glucose" and "cell." This dialog facilitated a broader range of student ideas by linking macroscopic and microscopic concepts and promoting discipline-specific vocabulary. These examples illustrate that while both dialogs prompted elaboration, the expanded model was more successful in guiding students to integrate scientific concepts and use scientific language, demonstrating the importance of dialog design in fostering scientific reasoning.

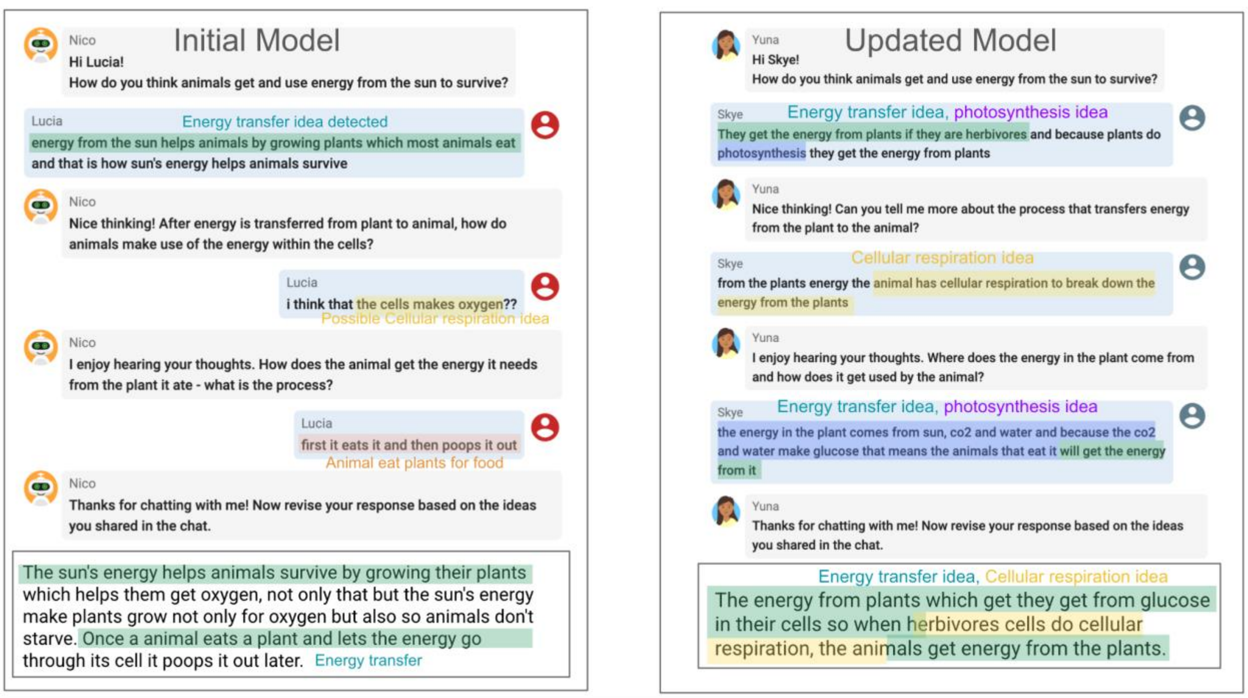

Figure 11. Student example using the dialog with the initial model (left, Lucia) and with the updated model (right, Skye)

5.4 Prompting Strategy: Prioritizing on Normative vs. Intuitive Ideas

Our RPP revised two design features at Iteration 4: guidance design (elicit-distinguish vs. distinguish-distinguish) and prompt strategy (guiding intuitive ideas first vs. guiding normative ideas first). The dialog involved two rounds of guidance. In Iteration 3, students received eliciting guidance in the first round and distinguishing guidance in the second round. Both rounds guided intuitive ideas first when students had two or more ideas. In Iteration 4, both rounds provided distinguishing guidance and prioritized normative ideas first for students who began with multiple ideas. Table 4 summarizes the design variations across Iterations 3 and 4.

Table 4. Design variations across Iterations 3 and 4.

Iteration	NLP model	Guidance design	Prompting strategy	Sample Size (N)
Iteration 3	Expanded (19 ideas)	Round 1: elicit Round 2: distinguish	Guide on intuitive ideas first when students have multiple ideas	Fall 2023, $N = 691$
Iteration 4	Expanded (19 ideas)	Round 1: distinguish Round 2: distinguish	Guide on normative ideas first when students have multiple ideas	Fall 2024, $N = 647$

Overall, student Knowledge Integration (KI) scores significantly increased from initial to revised explanations in both iterations ($\beta = 1.71$, $SE = 0.18$ $z = 9.46$, $p < 0.001$). CLMM analyses indicated that the combined effect of the prompting and guidance strategies used in Iteration 4 ($\beta = -1.04$, $SE = 0.38$, $z = -2.75$, $p = .006$). As Figure 12 shows, in Iteration 4, monolingual students in Iteration 4 showed greater gains in KI from initial to revised responses compared to multilingual students. I explored whether these effects could be explained by baseline differences in students' initial KI scores. Adding starting KI levels into the model did not yield significant interactions, suggesting that the observed differences were not solely attributable to students' initial understanding.

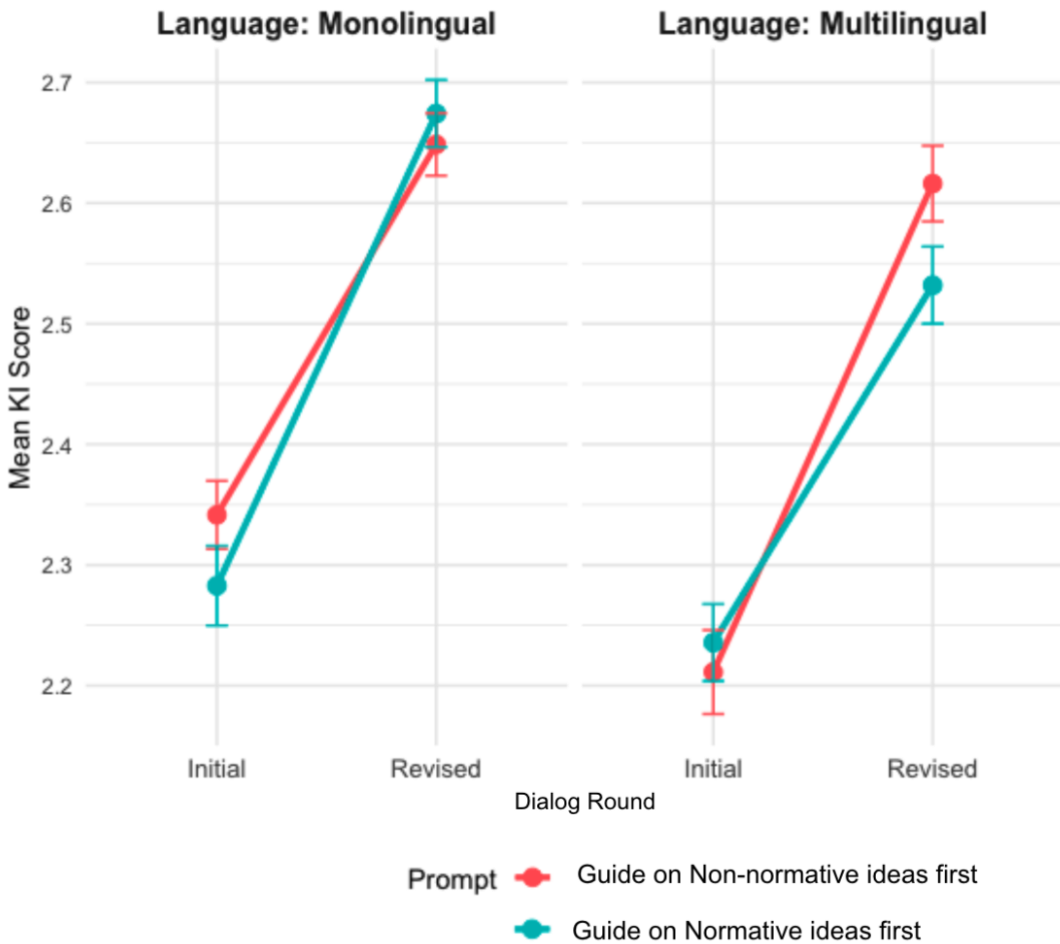

Figure 12. Changes in Mean KI scores by Dialog Round, Prompt strategy, and Language background

I then isolated the effects of each design feature, guidance type and prompt strategy, through targeted subgroup analyses.

To examine the independent impact of guidance type, I focused on students who began Round 1 with only one idea, across Iteration 3 (N = 267) and Iteration 4 (N = 272). For these students, prompt strategy was not applicable, as they had only one idea and thus received the specific prompt for that idea. Among these students, 67% initially expressed an intuitive idea, while 33% expressed a normative idea. Analyses using generalized linear models and zero-inflated models found no statistically significant differences between elicit and distinguish guidance types in either the total number of new ideas added ($p > .05$) or new normative ideas added ($p > .05$), even after controlling for language and grade. Students starting with only one idea may not have enough knowledge pieces to integrate another idea or distinguish multiple ideas, as shown in our data that over 93% of students with one idea are at lower KI levels (1 or 2) in both iterations. Detailed descriptive statistics are presented in Appendix Table 3.

Finally, to examine the isolated impact of prompt strategy, I focused on students who expressed two or more ideas in Round 2 across Iterations 3 (N = 234) and 4 (N = 126). Both iterations used distinguish guidance at round 2, differing only in their prompt strategies. Poisson and zero-inflated regression analyses indicated no significant differences between prioritizing intuitive versus normative ideas first in terms of total new ideas or new normative ideas added ($p > .05$). 57% of students in this subset started with at least one KI link at their initial explanation, suggesting that students might have exhausted most of their available ideas at the initial explanation and round 1, thus reducing the impact of prompt strategies in round 2. Descriptive statistics for this subset are provided in Appendix Table 4.

In summary, although our statistical analyses did not clearly identify the specific mechanisms by which guidance design and prompt strategies individually contributed to student idea incorporation and KI gain across dialog rounds, students demonstrated significant KI improvements in both iterations. Importantly, the design changes implemented in Iteration 4 appeared particularly effective for monolingual students, supporting deeper and more integrated understanding rather than superficial expansions of ideas. One possible explanation is that monolingual students may have found the distinguishing prompts linguistically more accessible, enabling them to engage more deeply with the cognitively demanding task of integrating ideas. However, this study did not directly measure linguistic processing demands, and alternative explanations, such as differences in prior science instruction or cultural ways of responding to assessment prompts, cannot be ruled out. Future research should explore additional support for multilingual learners to enhance their knowledge integration during adaptive dialogs.

6 Discussion

This study examined how iterative co-design with classroom teachers improved an AI adaptive dialog supporting middle school students' knowledge integration about photosynthesis and energy transfer. Over two years and four design iterations involving over 3,000 students, I investigated how different guidance types, idea detection rubrics, and prompting strategies affected students' ability to integrate scientific ideas. All four iterations showed significant KI gains after the dialog ($d = 0.30$ to 0.45), with controlled comparisons revealing which specific design features contributed to these improvements.

6.1 Three Design Principles for Adaptive Dialog Guidance

The controlled comparisons across iterations yielded three evidence-based design principles for adaptive dialog guidance grounded in the Knowledge Integration framework.

Principle 1: Use Eliciting Guidance for Idea Generation, Distinguishing Guidance for Idea Integration

The comparison of Iterations 1 and 2 revealed that two types of guidance served different pedagogical functions in the dialog. Eliciting guidance ("Can you tell me more about..." "What else happens...") supported idea generation: when students expressed no detectable ideas in their prior response, eliciting prompts marginally outperformed distinguishing prompts in generating new ideas ($IRR = 1.26$, $p = .078$). In contrast, distinguishing guidance ("How does X differ from Y?" "What is the process that...") supported idea integration: when students had already expressed at least one idea, distinguishing prompts were significantly more effective, leading to 60% more new ideas ($IRR = 1.60$, $p = .020$) and 73% more normative ideas ($IRR = 1.73$, $p = .029$) compared to eliciting prompts.

This finding extends prior research showing that guidance should adapt to learners' zone of proximal development (Vygotsky & Cole, 1978) and refines our understanding of how to adapt. When students have no entry point, they need invitation to share thinking. When students have expressed initial understanding, they benefit more from prompts that ask them to compare, explain mechanisms, or clarify relationships. This is the distinguishing process central to knowledge integration (Linn & Eylon, 2011). This aligns with Gerard, Holtman et al. 's (2025) finding that effective AI guidance creates productive space for students to pursue their next step in scientific reasoning, adapting to the nature of ideas students already hold.

Principle 2: Expand Idea Rubrics Through Teacher Expertise to Recognize Classroom Knowledge

The comparison of Iterations 2 and 3 demonstrated that teacher-informed rubric expansion significantly improved both idea detection and student learning. When three experienced teachers reviewed student dialogs, they identified six ideas frequently expressed by their students but absent from researchers' initial 13-idea rubric. These included entry-point ideas reflecting students' everyday observations ("sun helps plants survive"), curriculum-specific concepts from classroom instruction ("10% rule in energy pyramid"), and intuitive ideas blending scientific and personal understanding.

After expanding the rubric to 19 ideas and retraining the NLP model, students in Iteration 3 were 2.1 times more likely to improve their KI scores ($\beta = 0.75$, $p < .001$) compared to Iteration 2, and they added significantly more mechanistic ideas about energy transfer and animal cellular respiration ($p < .001$). Qualitative analysis showed that the expanded model enabled students like Skye to use scientific terminology ("glucose," "cellular respiration") and link concepts across organizational levels (from sun to plants to animals to cells), whereas students in earlier iterations tended to remain at macroscopic descriptions.

This finding challenges common NLP development practices where rubrics are created early and rarely revised (Kubsch et al., 2023). Our iterative approach of developing an initial rubric from standards and researcher analysis, then refining it based on teacher identification of ideas their students actually express, produced a model more responsive to the diverse ways students articulate scientific understanding. This matters particularly for students from culturally and linguistically diverse backgrounds, whose prior knowledge may not align with dominant framings in early-stage rubrics (Akgun & Greenhow, 2022; Morales-Doyle, 2017).

Principle 3: Prioritize Students' Most Sophisticated Ideas When Multiple Ideas Are Present

The comparison of Iterations 3 and 4 examined prompting strategy: which idea should the dialog respond to first when students express multiple ideas? Earlier iterations prioritized intuitive ideas (addressing misconceptions first), while Iteration 4 prioritized the most conceptually sophisticated normative idea. Combined with using distinguishing guidance in both rounds (rather than eliciting in Round 1), Iteration 4 supported significant KI gains, particularly for monolingual students ($\beta = -1.04$, $p = .006$ for the combined effect). Importantly, the Iteration 4 findings about prompting strategy should be interpreted cautiously. Because I changed both guidance type (both rounds distinguishing) and prompting strategy (normative-first) simultaneously, I cannot isolate which feature drove the observed benefits for monolingual students. Subgroup analyses examining each feature independently found no significant effects, suggesting that the combination of features, rather than either alone, may have supported integration. Future research using within-iteration randomization could disentangle these effects.

This finding extends previous research on conversational agents in two ways. First, whereas earlier dialog systems tailored prompts primarily to whether a student response matched an expected answer (Graesser et al., 2004), our approach tailored prompts to the types of ideas students expressed (normative vs. intuitive vs. broad entry-point ideas) and their conceptual sophistication. Second, our dialog adapted in real time based on what students wrote in prior rounds, rather than following pre-scripted conversation paths. This enabled the dialog to build on students' most developed thinking rather than defaulting to deficit-oriented sequences.

The shift reflects a move from deficit-oriented guidance (what students don't yet understand) to asset-oriented guidance (building on what students already know). When students had already articulated mechanistic understanding of energy transfer, the Iteration 4 dialog helped them deepen that understanding rather than redirecting to less-developed ideas. This aligns with KI's principle of treating all student ideas as resources (Linn & Eylon, 2011) and with culturally sustaining pedagogies that build on students' existing knowledge (Paris & Alim, 2017). However, multilingual students showed smaller gains than monolingual students in Iteration 4, suggesting that distinguishing prompts, while cognitively productive, may create additional linguistic demands. Future designs should incorporate explicit language scaffolding that makes connections between everyday and scientific language more visible (Lee & Buxton, 2011).

6.2 Teacher Co-Design as Essential to Dialog Effectiveness

Beyond these three design principles, a cross-cutting finding is that teacher co-design was essential at every phase of dialog development. Teachers did not simply implement researcher-designed tools; they actively shaped what ideas the model could detect, how guidance responded to those ideas, and how detected ideas connected to classroom instruction.

Teachers identified ideas that emerged from their class implementations (Iteration 3 rubric expansion), simplified language that created barriers for multilingual learners, and reframed prompts to align with classroom discourse patterns. Importantly, teachers brought knowledge of what students in their specific communities express and value, enabling the dialog to recognize ideas grounded in local contexts rather than only dominant norms. This challenges the common framing of AI as "scaling" teacher expertise by automating what expert teachers do. Our findings suggest a different relationship: AI can amplify teacher expertise by making adaptive guidance available to all students simultaneously, but only if teachers actively shape what that AI recognizes and how it responds. The iterative co-design process, not the NLP technology alone, produced dialogs effective for diverse learners.

6.3 Limitations

Several limitations constrain the claims I can make from this study.

First, the most significant limitation is that design changes were confounded with implementation timing and instructional context. Iterations 1 and 4 occurred in fall (start of year, no prior instruction), while Iterations 2 and 3 Spring occurred after a full year of science learning. Iteration 3 Spring uniquely involved three teachers implementing the dialog within a full instructional unit rather than as a standalone assessment. These contextual differences influenced both baseline KI scores and potential for improvement. I addressed this through controlled comparisons that held context constant where possible: comparing Iterations 1-2 (both using original model), 2-3 Spring (both end-of-year), and 3-4 Fall (both start-of-year). However, I could not fully isolate all potentially relevant factors. For example, schools varied in demographics (26-95% free/reduced lunch) and unmeasured aspects of instructional quality or culture. While I controlled for school as a random effect in statistical models, school-level differences may still partially explain observed effects.

Second, Iteration 4 changed both guidance type (distinguishing in both rounds) and prompting strategy (normative-first), making it impossible to isolate which change drove the observed differential benefits for monolingual versus multilingual students. A more rigorous design would have tested these features independently through within-iteration randomization, though this was not feasible given the RPP's implementation constraints.

Third, this study focused on one science topic (energy transfer in photosynthesis and ecosystems) embedded in one curriculum platform (WISE). The specific ideas students express, the guidance that supports integration, and the role of teacher expertise may differ for other science topics or in other learning environments. However, the three design principles identified here are grounded in Knowledge Integration theory and likely apply broadly, even if specific implementations differ.

Lastly, the KI scoring rubric rewards linking ideas within a single written explanation. While this measures students' ability to articulate connections, it may not fully capture deeper forms of knowledge consolidation that develop over longer timescales or manifest in application contexts. The dialog supported students in adding and connecting ideas during the immediate revision task, but I did not assess whether these integrated explanations transferred to new contexts or persisted over time (I will discuss how the dialogs work in instruction contexts in Chapter 3's study).

6.4 Implications for Design and Research

Despite these limitations, this study offers actionable guidance for researchers and designers developing AI-supported dialog systems for science education.

For dialog designers, the finding that eliciting guidance supports idea generation while distinguishing guidance supports idea integration suggests a practical implementation approach. This requires three components: an idea detection model achieving acceptable accuracy ($F1 \geq 0.68$), a guidance library with both prompt types, and conditional selection logic. The threshold can be simple: our findings suggest a binary rule (zero vs. one-or-more) is sufficient, though future work might test whether finer gradations (zero, one, two-or-more) provide additional benefit.

For NLP researchers, the rubric expansion finding challenges the common practice of creating rubrics early and freezing them. Model development should budget time for teacher review after initial deployment: I found that having teachers review 20 diverse student dialogs in 60-90 minute sessions identified emerging ideas that substantially improved both model performance and student outcomes. The practical question is creating efficient review processes that fit within development timelines: our cycle added approximately 3 months but yielded improvements in effect size from $d = 0.33$ to $d = 0.49$.

For research-practice partnerships, the effectiveness of our co-design process depended on granting teachers decision-making authority, not just soliciting feedback. This required resource allocation is about approximately 40 hours of teacher time per year beyond teaching duties, compensation, and structures ensuring rapid implementation of teacher-proposed changes. Partnerships that position teachers as advisors rather than co-designers risk producing tools that fail to recognize local ways of knowing or create barriers for diverse learners.

Finally, the finding that distinguishing guidance benefited monolingual more than multilingual students highlights that adaptivity around conceptual understanding must be paired with linguistic scaffolding. Future designs should consider providing sentence stems, example responses, or multimodal response options that maintain cognitive demand while reducing linguistic barriers, and making such supports available to all students rather than tracking multilingual learners into separate pathways.

6.5 Future Directions

This study focused on a benchmark assessment context (dialog without prior instruction). Chapter 3 demonstrated that dialog effectiveness increases when paired with teacher-designed classroom activities that consolidate dialog-elicited ideas. Future research should examine how dialogs and classroom instruction can be designed as integrated systems, with dialogs surfacing ideas and activities helping students distinguish and connect those ideas through extended sense-making.

Additionally, while this study examined text-based dialogs, many students may express scientific understanding more fully through multimodal communication like speech, drawing, gesture, demonstration. Future adaptive systems might incorporate multimodal input and response, potentially making scientific reasoning more accessible to students who struggle with written expression.

Finally, this study examined one model of RPP structure involving sustained collaboration over two years. As AI capabilities advance rapidly, research should examine what "lightweight" co-design structures might enable teachers to adapt AI tools to local contexts without requiring extensive researcher partnerships. For example, teacher-editable prompt libraries, tools for viewing what ideas the model detects in their own students' responses, or interfaces for adding locally relevant ideas to rubrics.

Chapter 3: Impact of science teachers' pedagogical goals in customizing an AI adaptive dialog to promote Knowledge Integration in a Research-Practice Partnership

Chapter summary

This study examines how middle school science teachers' pedagogical goals shaped customizations of an AI adaptive dialog embedded in a photosynthesis and ecosystems instruction. Through a sustained research-practice partnership, two teachers co-designed dialog guidance and classroom activities within a web-based inquiry science unit. The Knowledge Integration (KI) framework provided theoretical foundation, analytical tools (KI principles to categorize goals, KI processes to code instruction), and outcome measures (5-level scoring rubric). I asked: (1) How did teachers' pedagogical goals shape AI dialog guidance refinements? (2) How do teachers' pedagogical goals guide their enactment of their customized dialogs in the photosynthesis unit?

Methods: I identified teachers' pedagogical goals through open coding of pre-implementation meeting transcripts and post-implementation interviews, organizing themes using KI design principles, with member checking by teachers and RPP reviewers to validate interpretations. Two teacher case studies with 132 students (grades 6-8) examined learning across five timepoints (start-of-year benchmark assessment, pretest, embedded test, posttest, delayed posttest). For RQ1 (guidance refinements), I analyzed student dialog responses at two timepoints (start-of-year assessment and pretest) per teacher using generalized estimating equations (GEE) models to test whether revised guidance increased idea incorporation, and analyzed meeting transcripts for revision rationales aligned with identified goals. For RQ2 (instructional customizations), I used cumulative link mixed models (CLMMs) to examine KI score changes, GEE models to analyze idea frequency changes, and coded classroom observation fieldnotes using KI processes (elicit, discover, distinguish, connect) to characterize instruction.

Findings. Pedagogical goals functioned as organizing principles creating coherence across design decisions. Amy's goal of making connections visible led her to revise guidance toward process and mechanism questions and co-design ecosystem drawing activities emphasizing relationships of organisms. Her instruction emphasized distinguishing and connecting, producing sustained KI gains at delayed posttest ($\beta = 0.58$, $p = .052$) and significant growth in mechanistic ideas about energy flow in ecosystems (27.12% added animal cellular respiration after drawing activity, $p < .001$). Ben's goal of making science accessible for emergent multilinguals led him to simplify syntactic complexity and co-design scaffolded whole-class simulation labs. His instruction emphasized eliciting and discovering, producing KI gains ($\beta = 0.25$, $p = .021$) at posttest and strong mechanistic understanding (28.75% added photosynthesis reactants after labs, $p < .05$), though gains were not sustained at delayed posttest.

These findings identify pedagogical goals as the mechanism creating coherence between AI tool refinement and classroom instruction. When teachers refined guidance and designed activities aligned with their goals, students showed significant knowledge integration gains and added mechanistic ideas aligned with the instruction across different contexts. Design

recommendations include building AI systems with teacher editability, supporting teachers in articulating KI-aligned goals, and enabling flexible activity design emphasizing different KI processes.

1 Introduction

AI-supported dialog systems promise to make student thinking visible at scale and to provide timely, adaptive guidance that responds to learners' ideas (Gerard, Linn & Holtmann, 2024; Zhai, Yin, et al., 2020). Yet in classroom settings, these systems do not operate in isolation. Their pedagogical function is shaped by how teachers interpret, revise, and integrate them into ongoing instruction (Hutchins & Biswas, 2024). From a Knowledge Integration perspective (Linn & Gerard, 2024), what matters is not only whether an AI system can detect ideas accurately, but how its guidance is orchestrated with classroom activities to support students in eliciting, distinguishing, and connecting ideas in local contexts over time (Gerard et al., 2025; Nyaaba & Zhai, 2024). Without attention to teacher design and instructional coherence, AI dialogs risk functioning as isolated feedback tools rather than as integrated components of coherent learning environments where instruction builds on students' evolving understanding over time.

Research on co-design of educational technology has documented benefits when teachers participate in development processes (Bichler et al., 2021; Gerard, Spitulnik, & Linn, 2010; Penuel et al., 2007; Roschelle & Penuel, 2006; Severance et al., 2016). Teachers bring knowledge of student language patterns, classroom contexts, and local communities that researchers and developers often lack (Philip et al., 2022; Vossoughi & Bevan, 2014). Studies show that teacher-informed technology designs lead to better learning gains (Li, Liao, et al., 2024), better alignment with classroom practice (Svihla et al., 2015), and increased teacher agency (Jayathirtha et al., 2025; Xia et al., 2025). However, more research is needed on what form of expertise teachers contribute, how that expertise manifests in design decisions, and tracing the complete pathway from teacher input to student learning outcomes.

This study examined co-design outcomes in a multi-year research-practice partnership (RPP; Coburn & Penuel, 2016) that integrated middle school science teacher pedagogical knowledge into both AI dialog development and the design of classroom activities that incorporate the AI dialog. The AI dialog, embedded in a web-based inquiry science environment (WISE) unit, uses Natural Language Processing (NLP) to detect ideas in students' written explanations and provides adaptive prompts based on those ideas (Gerard et al., 2025; Li, Chang, et al., 2024). Rather than focusing only on co-design structures or participation models, this study traces how teachers' pedagogical goals surfaced in both co-design processes and co-design products, including the specific guidance revisions teachers proposed and the classroom activities they co-designed. Critically, I identify teachers' pedagogical goals, which mean their stated priorities for what students should learn and experience, as the organizing principle shaping these design decisions. Through detailed analysis of these two implementation cases across a full instructional unit (pretest, embedded test, posttest, and six-month delayed posttest), I asked:

- RQ1: How did teachers' pedagogical goals shape AI dialog guidance refinements?
- RQ2: What instructional pathways emerged when teachers enacted co-designed dialogs and classroom activities across the unit, and how did these pathways support students' knowledge integration?

This study contributes empirical evidence that pedagogical goals function as organizing principles that shape the AI dialog, web-based curriculum and instruction, and learning consequences of AI supported instruction within a web based inquiry science unit. I trace how these goals shaped both the guidance revisions teachers proposed (RQ1) and the classroom activities they co-designed (RQ2), creating different but productive instructional pathways that are visible in student learning patterns.

2 Theoretical Framework

The KI framework (Linn & Eylon, 2011) guides the RPP in design and refinement of the instructional materials; interpretation of observations and student work; and customization of classroom practices. KI builds on extensive research documenting that students come to science class with diverse ideas including normative, intuitive, and descriptive beliefs (diSessa, 1993; Inhelder & Piaget, 1958; Kuiper, 1994). Research on KI shows that effective instruction leverages the ideas students develop from varied experiences (Matuk & Linn, 2023). By supporting students in sorting out their own ideas and others they discover in science class, KI instruction reinforces students' own reasoning strategies and encourages students to increase the coherence of their views by discovering new ideas as well as refining their reasoning capabilities (Songer & Linn, 1991). KI research yielded *four instructional processes*: eliciting existing ideas, discovering new ideas, distinguishing among alternative ideas, and sorting out connections among ideas (Linn et al., 2023). KI research also surfaced *four design principles* (Kali, 2006): (1) making science accessible to each learner with relevant science examples (Linn & Hsi, 2000); (2) making student thinking visible by posing explanation activities that require linking ideas with evidence (Ritchhart & Perkins, 2008); (3) helping students collaborate to learn from others (e.g. Matuk & Linn, 2023); and (4) promoting autonomy by guiding reasoning rather than providing answers (Davis, 2003). The RPP refines and applies these principles across curriculum design, software design, applications of AI, teacher professional learning, and evaluation design.

KI has three primary roles in this study (Table 1). First, KI design principles guided the design of the instructional materials and provided a lens for understanding how teachers' pedagogical goals shaped their decisions as they customized the materials. I use the term "pedagogical goals" to describe the organizing principles that teachers use to explain their decisions. Thus, when teachers articulated priorities such as "helping students see connections" or "making content accessible for emergent bilinguals," I linked these goals to associated KI design principles embedded in the curriculum. Second, KI processes provided analytic categories for coding teacher practices during classroom practices. I identified when teachers' practices encouraged students to elicit, discover, distinguish, or sort out their ideas. I coded teacher customizations of instruction to the KI processes. Third, the five-level KI scoring rubric assessed student learning outcomes by measuring integrated understanding and rewarding students for creating valid links among ideas rather than penalizing intuitions (Ryoo & Linn, 2015).

Table 1. Features of the Knowledge Integration Framework

KI Feature	Definition	Application	Results
KI Design Principles (Kali, 2006)	Make science accessible, Make thinking visible, Help students learn from others, Promote autonomy	Analytic codes for teachers' pedagogical goals in meeting and interview transcripts, unit customizations	Methods 3.5, Findings throughout
KI Processes for integrating ideas (Linn & Eylon, 2011)	Elicit your ideas, Discover new ideas, Distinguish among ideas, Sort and connect ideas	Analytic codes for observations of classroom instruction and interpretation of field notes.	Methods 3.6, Findings 4.2
KI Questions; Scoring Rubric (5 levels)	Explain complex topic; reward linking ideas with evidence (1=off-topic to 5=multiple links)	Indicator of student learning on pretest, postest, and during dialog	Methods Table 3; Findings

Our Research Practice Partnership (RPP) brings together researchers and practitioners in a long-term, mutually beneficial collaboration to improve educational practice by addressing shared problems (e.g., Coburn & Penuel, 2016). The RPP spans more than two decades beginning with the Computer as Laboratory Partner heralding the first uses of personal computers in classrooms (Linn & Hsi, 2000). With the advent of the Internet, the RPP designed the Web-based Inquiry Science Environment (WISE) and refined it using design studies in which teachers piloted web-based inquiry units, surfaced patterns in students' diverse and conflicting ideas, and jointly negotiated revisions to curriculum features and guidance with researchers across multiple years (Linn, Clark, & Slotta, 2003). Central to this partnership is the practice of curriculum customization: teachers' iterative adaptation of instructional materials to align with their own pedagogical priorities and with the needs, ideas, and experiences of their specific students and communities (Gerard, Spitulnik, & Linn, 2010). Empirical work explores the design of professional learning experiences that support teachers to use logged student data to interpret patterns in student thinking and make targeted instructional adjustments. The students of teachers who participate in these professional learning opportunities demonstrated deeper knowledge integration than students whose teachers received shorter, less contextually grounded support (Gerard et al., 2011). The RPP connects analysis of student work and customization of instruction as a designed aspect of professional learning (Bichler et al., 2021; Bradford, Bichler, & Linn, 2021; Gerard, Bradford, & Linn, 2022).

Effective customization depends on supporting teachers to notice their students' ideas (Luna, 2018; Jacobs, Lamb, & Philipp, 2010). The RPP created activities where teachers review their students' logged responses, interpret what those ideas reveal about the instruction, and experiment with customizations to respond to student needs (Bradford et al., 2021). The Curriculum Visualizer makes explicit the design principles implemented in the instructional materials (Gerard, Bradford, & Linn, 2022). Pre-implementation meetings enable teachers to develop and act on their interpretations of student ideas as they plan instruction (Billings & Linn, 2024).

As the partnership's work has extended to AI-supported dialogs, teacher noticing has taken on new significance. Besides analyzing student work, teachers can review the character and frequency of student ideas captured in NLP idea-detection models. In one study, teachers identified ideas that were missed during the initial model design, enabling the designers to refine the model and detect significantly more ideas in subsequent dialogs (Li, Liao, et al., 2024).

The present study examines how the RPP supported two teachers to refine their pedagogical goals and apply these goals in customizing a photosynthesis unit featuring an AI adaptive dialog and accompanying classroom practices. Specifically, I investigate how professional learning activities enabled teachers to articulate KI-aligned pedagogical goals, how reviewing student dialog data helped teachers notice specific gaps in student understanding, and how these noticings drove systematic revisions to both dialog guidance and classroom practices. The KI framework provides both the instructional foundation and the analytical lens for examining these processes.

3 Methods

3.1 Study Context: WISE unit and AI dialog

The Energy Story dialog was embedded within a 4-6 week WISE unit on photosynthesis and ecosystems, anchored by a school garden phenomenon (Figure 1). Students learned about photosynthesis and cellular respiration through interactive simulations, then applied this knowledge to plan for growing fruits and vegetables in a school garden (Lesson 3: Garden Plan Challenge). Later lessons explored food access inequities in their neighborhoods and designed actionable responses to food apartheid (Lesson 5: Food Justice). Teachers could customize the WISE unit sequence, add supplementary activities, and modify pacing based on their classroom needs. This flexibility recognizes that effective implementation requires local adaptation (Penuel et al., 2007).

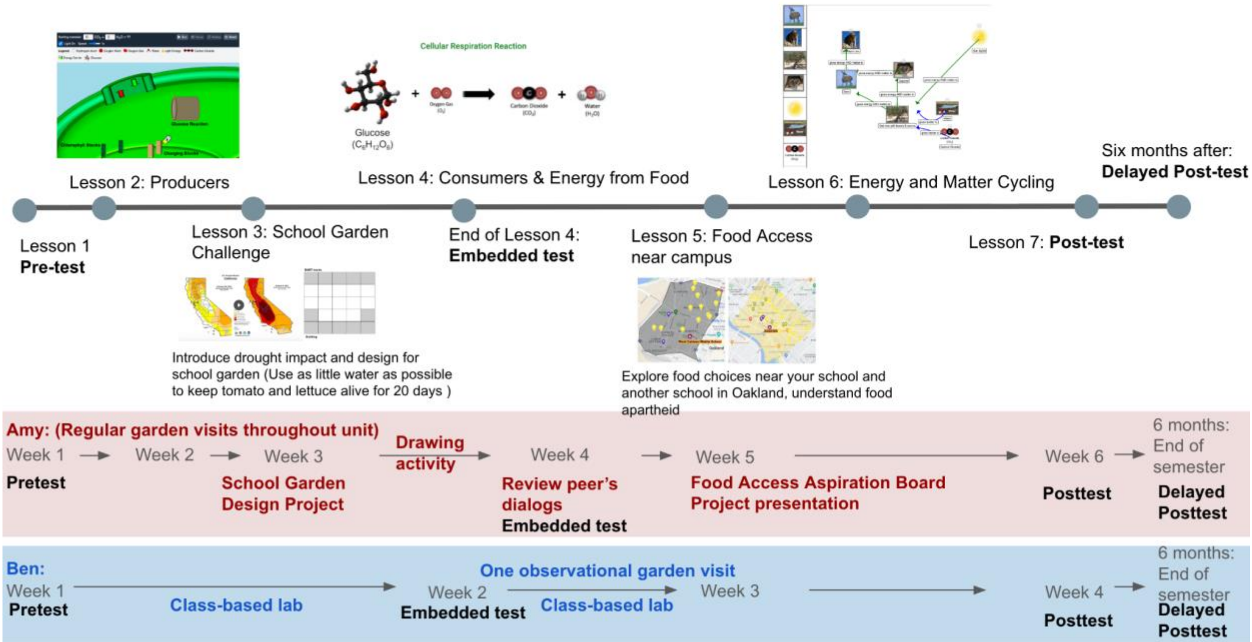

Figure 1. The WISE Photosynthesis and Ecosystems Unit

Students engaged with an AI adaptive dialog (Figure 2) embedded in the Web-based Inquiry Science Environment (WISE) platform as part of a photosynthesis and ecosystems unit. The dialog appeared at four timepoints: pretest (before instruction), embedded test (after Lesson 4), posttest (after Lesson 7), and delayed posttest (6 months later). This repeated-measures design allowed us to track students' knowledge integration development across the unit and assess retention over time. The Energy Story dialog asked students to explain “How does energy from the sun help animals to survive?” Students wrote an initial explanation. The KI scoring model uses the rubric to immediately assign a KI score to their initial explanation, and the NLP idea detection model immediately identifies the ideas in the explanation. Students receive two rounds of adaptive guidance based on ideas detected by the NLP model. The dialog guidance was originally developed through prior RPP design cycles (Chapter 2). In this study, Amy and Ben further refined the guidance based on their knowledge of their specific students and classroom contexts. After responding to both prompts, students revised their initial explanation. Finally, the KI scoring model automatically assigns a KI score to the revised explanation.

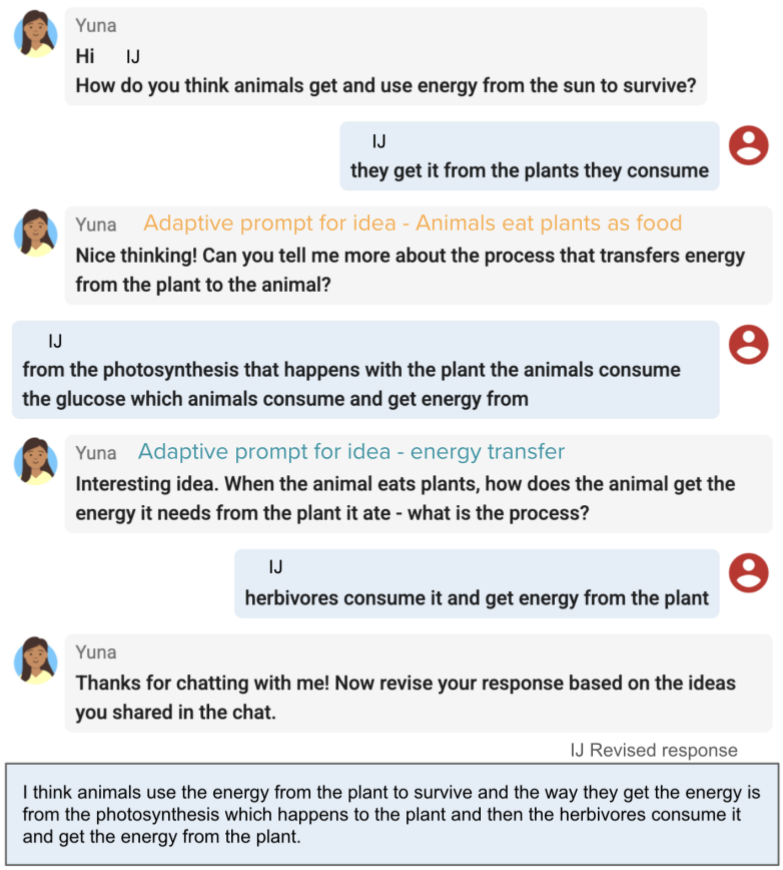

Figure 2. The Energy Story AI dialog between Yuna (a thought-buddy) and IJ (a 7th-grade student)

The KI scoring model in the Energy Story AI dialog was developed and validated in previous studies (Holtmann et al., 2023; Riordan, Cahill, et al., 2020; Ryoo & Linn, 2015; Li, 2025). Because the previously developed KI scoring model (See the 5-level rubric in Table 3) had a high agreement with human raters (QWK = 0.809), in this study, I used it to score students' initial revised explanations. Details for the model development approach can be found in (Riordan, Cahill, et al., 2020).

Table 3. The Energy Story AI dialog KI level description

Energy Story KI question: *How does energy from the sun help animals to survive?*		
KI	KI level description	Student Response Example
1	Responses are off-topic.	"I don't know "If you end this conversation then you are a preset and not a real AI"
2	**Vague**: Non-normative or scientifically invalid ideas	"Animals get vitamin D from the sun to strengthen their bone"

3	**Partial Link**: Normative ideas about the energy from the sun to plants to animals but did not include any normative links	"The sun helps animals survive by the sun helps plants make food so other animals can eat and then get the energy from the sun"
4	**One Full link**: Students included one valid reasoning of normative ideas	"The sun helps animals survive by the plants absorb the energy and perform photosynthesis. After that the energy is used for food in the plant and when an animal eats hit the energy is transferred from the plant to the animal."
5	**Complex links**: Students included two valid reasoning of normative ideas	"the plants use their chloroplast to change the light energy from the sun into useable energy. when the animal eats the plant, it gets the plant's energy, which it turns into glucose. the glucose gets sent to the mitochondria, which turns it back into energy"

The NLP idea detection model uses a multi-label token classification approach to detect up to 19 distinct ideas in student responses (Figure 3). The model was trained on 1,500+ annotated student responses and achieved a micro-averaged F1-score of 0.73 (Li, Liao, et al., 2024), acceptable for classroom deployment (Schulz et al., 2018).

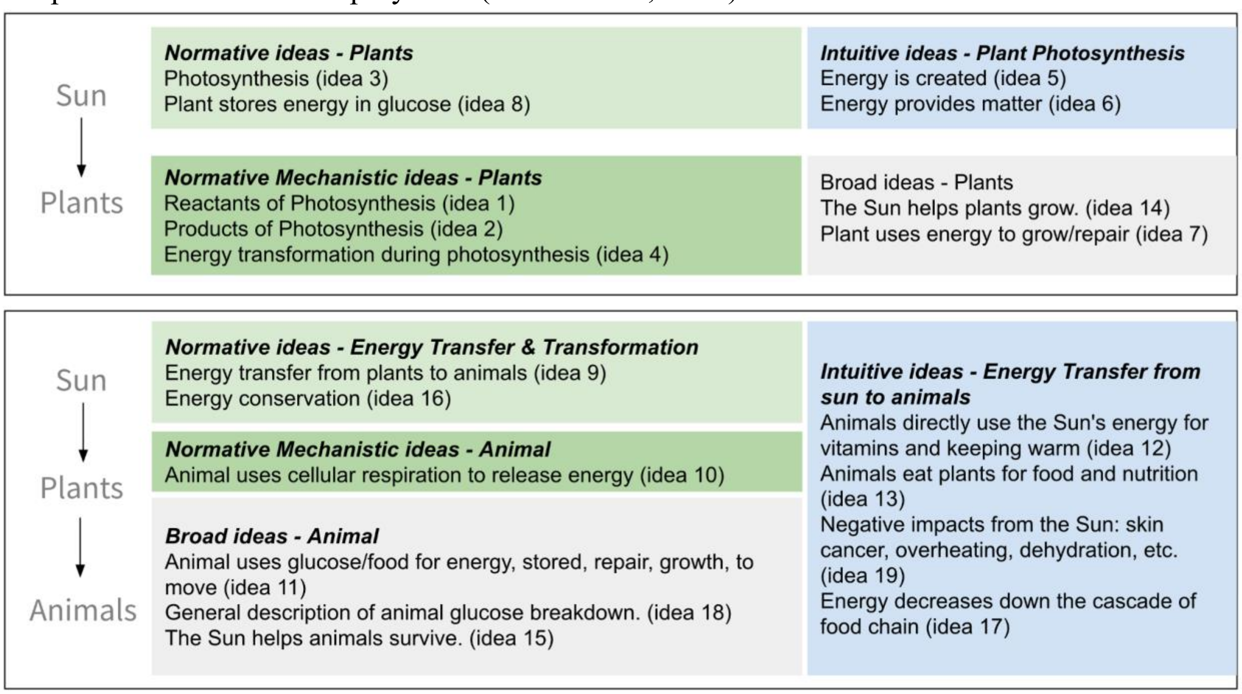

Figure 3. Idea rubric for the Energy Story AI dialog

3.2 RPP Participants

The RPP, as noted above, has evolved over decades, bringing together classroom teachers, learning scientists (including faculty, postdocs, and students), computer scientists, and curriculum designers to strengthen science instruction by leveraging technology and refining the KI pedagogy (Billings & Linn, 2024; Linn, Donnelly-Hermosillo, & Gerard, 2023). Starting with web-based curriculum materials, the partnership has incorporated real time graphs, visualizations, concept maps, collaborative tools, and AI innovations. The university members of the partnership meet weekly to critique advances in curriculum design, review research findings, monitor collaborations between learning scientists and classroom teachers, and plan workshops.

This study is a collaboration between two middle school science teachers, learning scientists and computer scientists. The teachers, Amy (pseudonym) has 10+ years of experience teaching 7th-8th grade science while Ben (pseudonym), is a second-year 6th grade teacher who joined the RPP during his first year of teaching (Table 2). Both teach in urban public schools in the western United States serving predominantly multilingual students. Amy's school serves 93% Hispanic students, with 86% qualifying for free/reduced lunch and 80% speaking languages other than English at home. Ben's school serves 90% Hispanic students, with 95% qualifying for free/reduced lunch and 85% speaking languages other than English at home.

The teachers differed in their experience with KI and WISE, their classroom practices, and their knowledge of their school context (Table 2). Amy had deep familiarity with her students' background, and established classroom practices. She drew on her established practices to customize new materials. Ben, in his second year of teaching, recognized his students' limited prior science instruction, skepticism about the value of science in their lives, and desire to understand science lessons. He often revised textbook questions for his emergent bilingual students.

Table 2. Teacher Experience and Approach

Dimension	**Amy**	**Ben**
Teaching Background		
Teaching Experience	10+ years (7th-8th grade science)	2nd year (6th grade science)
Teaching preparation	2 years MA program, introduced to KI and WISE during the program	1 year MA program, introduced to KI and WISE in courses and MA projet
RPP involvement	7 years; enacted 35 WISE units	1 year; mentored for MA project
Classroom Context		
Student sample	N = 69 (3 classes, ~20/class)	N = 63 (2 classes, ~30/class)
School demographics	93% Hispanic; 86% free/reduced lunch; 80% multilingual	90% Hispanic; 95% free/reduced lunch; 85% multilingual

Implementation		
Photosynthesis and Ecosystems Unit duration	6 weeks (~3-4 periods/week); 2 weeks on photosynthesis (including school garden project), 2 weeks on cellular respiration, 1 week on ecosystem, 1 week on food justice.	4 weeks (~2-3 periods/week); 2 weeks on photosynthesis (including school garden project), 1 week on cellular respiration, 1 week on ecosystem and food justice.
Prior science instruction	Students started science in 6th grade; studied Ecology unit before the Photosynthesis unit	Students starting science in 6th grade.
Garden engagement	Regular school garden visits and hands-on gardening work throughout unit in collaboration with school gardening teacher	One observational garden visit during the photosynthesis lesson
Pedagogical Approach		
Noticing of student ideas	Students had mix of descriptive and mechanistic ideas; sought opportunities to link ideas	Emergent bilingual students used a wide range of examples to describe their science ideas; questioned the relevance of science to their lives.
KI pedagogical goal	Making connections visible	Making science accessible
Customization of instruction	Designed an ecosystem drawing activity to elicit links among ideas.	Designed a whole-class simulation lab to illustrate the connections of science to everyday life.
Classroom practices	Used exit tickets and homework drawing activities to help students visualize science	Used AI and other tools to make textbook questions accessible to his emergent bilingual students.

3.3 Researcher Positionality

The first author is a PhD candidate focused on learning analytics and technology design. Her limited U.S. K-12 classroom teaching experience due to her international background led her to position herself as a collaborator learning from teachers' pedagogical expertise rather than as an external expert. Throughout the study, she maintained a stance of inquiry, asking teachers to explain their reasoning, share their observations of student work, and propose design changes based on their classroom knowledge. This positioning was essential to creating space for teachers' expertise to shape both the AI dialog and the research itself. The second author serves as a mentor in the RPP. She meets regularly with each participant and leads classes that review technological innovations and grapple with interpretations of findings from RPP activities.

3.4 RPP instructional customization process

RPP customizations occurred in iterative cycles across the broader RPP over multiple years. This study focuses on RPP customizations with Amy and Ben during a two-year period within this timeframe, with each cycle including four phases (Figure 4): (1) Workshops (twice yearly) where teachers reviewed web-based curriculum materials and their students' dialog data from the previous semester, discussed KI pedagogy and how it was enacted in the curriculum, and identified potential customizations to the AI dialog guidance and unit activities; (2) Pre-implementation meetings (2-4 weeks before instruction) where teachers and the first author discussed specific dialog guidance revisions, unit adaptations, and locally relevant classroom practices; (3) In-the-moment collaboration during implementation, including conversations between class periods and, when needed, rapid revisions to the web-based curriculum; and (4) Post-implementation meeting where researchers and teachers collaborative review of student work after implementation to identify needed revisions.

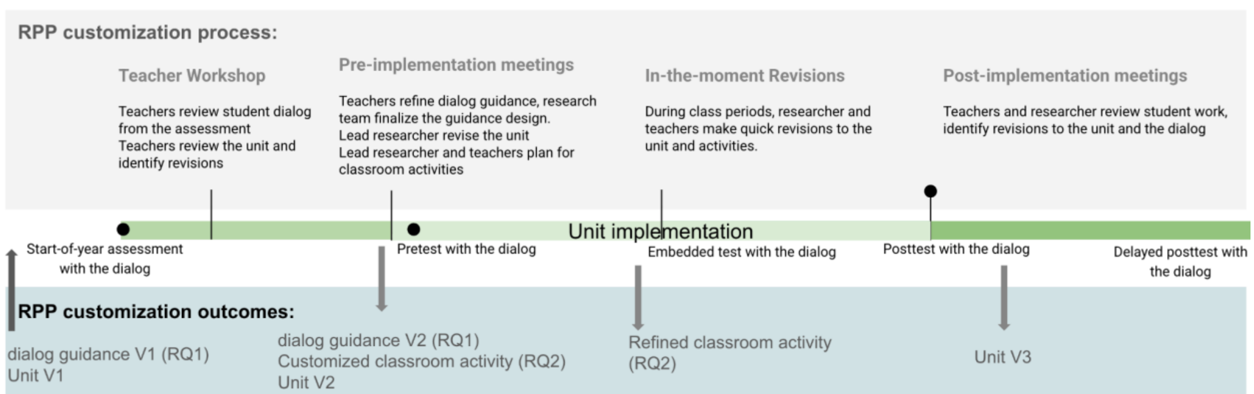

Figure 4. RPP customization process and products before, across, and after the unit implementation

Within this RPP, customization followed a structured process designed to cultivate and sustain teacher agency by positioning teachers as the primary interpreters of their students' ideas and the primary agents shaping design decisions (Luna, 2018; Gerard, Bichler, et al., 2022). Teachers reviewed their students' responses to the AI dialog featuring the original RPP-designed guidance, identified areas for revision based on their observations of student work and classroom patterns, and proposed specific changes grounded in their knowledge of their students' ideas and community contexts. The first author incorporated these revisions in consultation with the broader RPP team, and teachers reviewed revised guidance before implementation. Specific examples of guidance revisions and their pedagogical rationales are presented in the RQ1 findings (Section 5.1). Teachers also proposed classroom practices to help students integrate ideas elicited by the dialog; the first author provided technical support (e.g., creating digital resources, modifying WISE unit structure) and ensured alignment with KI design principles, while teachers made decisions about activity content and instructional approach. This structure where teacher noticing of student ideas drove customization decisions and researcher support enabled their enactment distinguishes this RPP's approach from researcher-led design cycles and reflects the partnership's broader commitment to promoting teacher autonomy as a condition for responsive, localized science instruction (Linn & Hsi, 2000).

Sustaining this structure required ongoing, embedded collaboration. The first author was present in classrooms throughout implementation, observing instruction, taking fieldnotes, and having informal conversations with teachers about how the dialog and practices were working. This embedded presence supported responsive collaboration, allowing teachers to propose adjustments and the researcher to document the reasoning behind design decisions.

3.5 Data sources and analysis

This study employed a multiple case study approach (Yin, 2018) to examine how pedagogical goals shaped customization in two classroom contexts. Data sources included pre-implementation meeting transcripts (two per teacher), post-implementation interview transcripts (one per teacher), classroom observation fieldnotes (present for 90%+ of sessions), and logged student dialog data across multiple timepoints.

Identifying pedagogical goals. I analyzed pre-implementation meeting transcripts and post-implementation interviews to identify each teacher's pedagogical goals: their stated priorities for what students should learn, how they typically supported student learning, and why particular outcomes mattered. Goals were not predetermined but emerged through iterative analysis of teachers' statements and design decisions, and reflected practices teachers had developed through RPP courses and workshops grounded in KI. Through open coding (Strauss & Corbin, 1998), I identified recurring themes, then organized them using the four KI design principles as an analytical framework (Kali, 2006), because teachers had developed familiarity with these principles through the RPP and because KI research has shown these principles to be consequential for inquiry-based science instruction (Linn & Eylon, 2011). To confirm our interpretations, I conducted member checking with three reviewers (Lincoln & Guba, 1985). Two teachers each received a written summary of their identified pedagogical goals and was asked: “Does this accurately capture what you were trying to accomplish with your students? What would you add, change, or clarify?” Additionally, a learning scientist from the RPP team reviewed both goal characterizations for alignment with KI design principles. All three reviewers confirmed our interpretations.

RQ1 analysis. To examine how teachers' pedagogical goals shaped guidance revisions, I analyzed pre-implementation meeting notes, revised guidance texts, and student response patterns. I conducted two statistical analyses per teacher using Poisson GEE models accounting for repeated measures within students. Analysis 1 compared student responses to revised versus original guidance for each teacher's revised items (Amy: 6 items, 140 observations; Ben: 4 items, 73 observations), testing whether students added more ideas when receiving revised guidance. Analysis 2 tested whether improvement from start-of-year assessment to pretest (no instruction yet) was larger for ideas with revised guidance compared to ideas with unchanged guidance, using all student guidance-response observations (Amy: 276 observations; Ben: 252 observations). Both analyses used Timepoint (Start-of-year Assessment vs. Unit Pretest) as the predictor. I tested whether to include covariates for guidance item, dialog round, student language background, and idea category; none significantly improved model fit or were significant predictors ($p > .10$), and including guidance items as a covariate was problematic given small sample sizes per item ($N = 3–21$ observations). Final models therefore included only Timepoint as a predictor.

Because all students received original guidance at the Start-of-year Assessment and revised guidance at the Unit Pretest, timepoint and guidance version are confounded. I cannot statistically isolate guidance effects from three months of intervening science instruction. I address this limitation through convergent evidence: systematic revision patterns, aggregate statistical improvements, differential improvement for revised versus unchanged items (Analysis 2), and qualitative examples of student elaboration.

RQ2 analysis. For RQ2 (examining instructional customizations), I collected dialog data across four timepoints: pretest, embedded test (after Lesson 4), posttest (after Lesson 7), and delayed posttest (6 months later). All student explanations received Knowledge Integration (KI) scores and detected idea labels from NLP models in all responses. I used cumulative link mixed models (CLMMs) to examine KI score changes across timepoints, appropriate for ordinal data, with fixed effects for timepoint and student demographics and random intercepts for student and school. I also examined idea frequency changes using generalized estimating equations (GEE) with binomial distribution. Qualitatively, two researchers (the first author and a pre-service teacher researcher) coded teacher instruction during classroom practices using deductive coding (Miles, Huberman & Saldaña, 2014) with KI processes (elicit, discover, distinguish, reflect/connect; Linn & Eylon, 2011) from classroom observation fieldnotes (present for 90%+ of sessions), and in-the-moment between periods notes. I used KI processes as analytic codes because these categories capture what kind of thinking teachers prompted students to engage in, distinct from their broader pedagogical goals about learning environments. I analyzed whether students' learning patterns reflected the KI processes emphasized in teacher-designed practices and whether outcomes aligned with teachers' pedagogical goals.

I note that the study design does not allow us to isolate independent effects of dialog versus classroom practices, as students experienced both integrated into instruction. Therefore, findings describe the combined effectiveness of customized dialogs and practices rather than attributing gains to one component alone.

4 Results

I analyze the teachers' pedagogical goals. I investigate how the goals impact the customization of the units and classroom practices.

4.1 Teachers' Pedagogical Goals

Amy's pedagogical goals centered on two priorities, both aligning with KI design principles. First, she emphasized making thinking visible by helping students see relationships among ecosystem components rather than treating them as isolated facts. She wanted students to make connections visible, what she called “understanding the whole chain” and “seeing how things connect.” When discussing student responses that listed disconnected facts, she noted: “They know plants need sun, animals eat plants, but they don't connect it. I need them to see it's not three separate things, it's one system where energy moves through.” Second, she prioritized making science accessible through personal relevance, connecting scientific concepts to students' lived experiences in their urban community, particularly through their school garden: “A lot of curriculum is about rainforests or oceans or places they've never been. But we have this garden right here. If I can help them see the ecosystem science in their own space, it becomes real.”

Ben's pedagogical goals also centered on two priorities, both focused on making science accessible. First, he emphasized making science accessible through scaffolded discovery, ensuring complex concepts were comprehensible to students with limited prior science instruction. He focused on building shared understanding through careful scaffolding before independent work: “A lot of them haven't had any science instruction up till now. I can't assume they know what photosynthesis means. I have to build it step by step.” This reflects the KI principle of seeking an appropriate level of analysis so students can restructure and develop more cohesive ideas (Linn et al., 2004). Second, he prioritized making science accessible through linguistic clarity, removing language barriers for emergent bilinguals: “If the sentence is too long with too many clauses, they just shut down. I need to break it down so they can actually process what we're asking.”

While Amy articulated two pedagogical goals, this study traces primarily the connections goal because it most directly shaped both dialog guidance refinements and activity design. Amy's personal relevance goal (using the school garden) manifested in the context for her practices (drawing their own garden rather than a generic ecosystem) but operated as a complementary priority supporting engagement rather than organizing design decisions across multiple levels. Ben similarly articulated two goals (scaffolded discovery and linguistic clarity), and this study traces both because they shaped distinct design decisions: scaffolded discovery shaped activity structure (whole-class labs), while linguistic clarity shaped guidance revisions.

I analyzed how these pedagogical goals shaped dialog guidance refinements (RQ1) and instructional customizations (RQ2). I examined how teachers’ design decisions aligned with their stated goals and how student outcomes reflected goal-aligned emphases.

4.2 RQ1: Dialog Guidance Refinements in Two Cases

Teacher customization measurably improved dialog guidance effectiveness. Each teacher's revisions aligned with distinct pedagogical goals that organized their customizations across multiple design levels. Amy's revisions, driven by her goal of helping students make connections visible, shifted guidance from categorization questions to process questions and were associated with more mechanistic student reasoning. Ben's revisions, driven by his goal of making science accessible through linguistic clarity, simplified complex sentence structures that had created comprehension barriers for emergent bilinguals. Both teachers' refinements produced statistically significant improvements in student responses. For each teacher, I present systematic revision patterns aligned with their pedagogical goals, aggregate statistical results, and student examples illustrating how revisions supported elaboration.

4.2.1 Systematic Revision Patterns Across Teachers

Both teachers systematically revised dialog guidance to align with their pedagogical goals, but applied distinct strategies shaped by their students' needs and contexts (Table 4). Amy, teaching 69 students across 7th-8th grades (80% multilingual), revised six guidance items to make energy connections visible. Ben, teaching 63 sixth-grade students (85% multilingual emergent bilinguals with limited prior science instruction), revised four items to make science linguistically accessible.

Table 4. Teachers' Goal-Aligned Guidance Revisions

Teachers' Systematic Revision Strategies: **Amy**: Shifted from categorization/listing to process explanation; replaced abstract process-naming with concrete mechanisms; emphasized dynamic energy and matter movement through systems **Ben**: Reduced compound questions; replaced abstract nouns with concrete action verbs; simplified syntax and academic vocabulary to everyday language; added spatial context for grounding			
Idea description	**Original Guidance**	**Amy's Revised Guidance**	**Ben's Revised Guidance**
Intuitive and Precursor Ideas			
Animals eat plants for food and nutrition (Idea 12)	"Interesting idea! **What type of energy** do animals get from plants?"	"Interesting ideas! **What happens to the energy** after the animal eats the plant?"	(Not Revised due to limited time in the pre-implementation meeting)
Sun helps plants survive (Idea 3)	"Can you tell me more about how the Sun helps plants survive? What is the process?"	"How does the sun help plants grow **and provide energy to** animals?"	"**How does the plant use** the energy from the sun?"
Sun helps animals (Idea 11)	"I see, you are saying that animals need the sun to survive. **Can you give me an example** of how the sun helps animals?"	"Interesting! **How does the sun help the animal live?**"	(Not Revised due to limited time in the pre-implementation meeting)
Animals directly use sun's energy (Idea 11a)	"The sun helps animals **in lots of ways**! How do you think the sun helps animals get energy from food?"	"Good progress! **How does the animal use the energy from the sun?**"	"**What do animals need to do to get energy**, other than absorbing sunlight?"
Normative Ideas			
Energy transfer (Idea 12b)	"Nice thinking! Can you tell me more about the **process** that transfers energy from the plant to the animal?"	"Tell me more! **How does eating the plant give the animal energy?**"	"**How does the energy** from the plant **get used inside the animal**?"
Photosynthesis (Idea 3b)	"Nice thinking about photosynthesis. How are the **products** of photosynthesis **useful** for animals?"	"Good progress! How does the energy from photosynthesis make its way to animals?"	"How do the **products** of photosynthesis **help** animals **live**?"

Consistent with Amy's goal of making connections visible through connecting ideas, in our pre-implementation meetings, she revised guidance to emphasize mechanisms and relationships rather than categorization. Her revisions focused on helping students articulate how energy moves through ecosystems. As she emphasized in the pre-meetings, "I want them to understand the whole chain, not just that plants make food. How does the sun's energy actually get to the cow? That's what understanding ecosystems means, seeing how it all flows together."

Across all six revisions (Table 4), she shifted from categorization questions to process questions (e.g., “What type of energy?” became “What happens to the energy?”), from static language to dynamic flow language (e.g., “How are products useful?” became “How does energy make its way?”), and from compound prompts to focused causal questions. These revisions reflected her pedagogical priority: making energy transfer relationships explicitly visible rather than asking students to list isolated facts.

Consistent with Ben's goal of making science accessible through linguistic clarity, he revised guidance to remove syntactic complexity that created comprehension barriers for emergent bilinguals. His revisions focused on simplifying sentence structures, reducing clause embedding, and placing key concepts early in sentences. Across his four revisions (Table 4), he reduced embedded clauses (e.g., compound questions became single questions), replaced abstract nouns with concrete action verbs (e.g., “sources of energy” became “need to do”), and used everyday language instead of academic vocabulary (e.g., “process” became “get used inside the animal”). As he explained in our pre-implementation meeting regarding original guidance with compound questions: “My students have to hold all of that in working memory while figuring out what we’re asking. My ELs can’t do that. By the time they get to ‘what is the process,’ they’ve lost track of the subject.”

Table 4 illustrates how both teachers revised guidance for overlapping idea categories but applied different strategies. For instance, both revised guidance for Idea 3 (sun helps plants survive), starting with the same problematic compound question. Amy's revision emphasized the mechanism of energy transfer in the food chain from the Sun to plants then to animals ("How does the sun help plants grow and provide energy to animals?"), focusing on making the process visible. Ben's revision simplified the syntax while retaining the core question about energy use ("How does the plant use the energy?"). Similarly, for Idea 3b (photosynthesis), Amy shifted to dynamic flow language ("make its way"), while Ben shifted to everyday language ("help...live"). These parallel revisions demonstrate how pedagogical goals shaped specific design decisions, even when addressing the same underlying student ideas.

4.2.2 Amy’s Student Outcomes

To test whether Amy's connection-focused revisions (Table 2) improved student responses, I conducted two GEE analyses.

Analysis 1: Six revised items. Filtering to instances where students received at least one of Amy's six revised prompts yielded 140 guidance-response observations (61 from Start-of-year Assessment, 79 from Unit Pretest). I examined two outcomes: (1) new ideas of any type (normative, intuitive, or precursor), and (2) new normative ideas specifically. Students added significantly more new ideas receiving revised guidance at Pretest ($M = 0.62$, $SD = 0.80$) compared to original guidance at Start-of-year Assessment ($M = 0.35$, $SD = 0.62$) ($\beta = 0.58$, $SE = 0.19$, $z = 3.05$, $p = .002$; IRR = 1.79, 95% CI [1.22, 2.62]). Students added significantly more normative ideas when receiving revised guidance at Pretest ($M = 0.48$, $SD = 0.71$) compared to original guidance at Start-of-year Assessment ($M = 0.22$, $SD = 0.49$) ($\beta = 0.78$, $SE = 0.18$, $z = 4.33$, $p < .001$; IRR = 2.18, 95% CI [1.53, 3.11]). Students added more than twice as many normative ideas with Amy's connection-focused revisions. Dialog round was not a significant predictor ($p = .52$), indicating guidance effectiveness did not vary by position in the dialog.

Analysis 2: Revised vs. unchanged comparison. Using all 276 guidance-response observations (69 students × 2 timepoints × 2 rounds), the Timepoint × IsRevised interaction was significant ($\beta = 0.52$, $SE = 0.21$, $z = 2.48$, $p = .013$), indicating improvement was significantly larger for revised items (IRR = 2.18) compared to unchanged items (IRR = 1.37). This differential improvement (59% more growth for revised items) supports the inference that Amy's specific revisions, rather than general time or learning effects alone, contributed to enhanced mechanistic reasoning.

4.2.3 Amy's Student Cases

Student examples illustrate how Amy's connection-focused revisions supported mechanistic reasoning. Figure 5 compares two students' dialog experiences with original versus revised guidance for the same idea (Idea 3: sun helps plants survive).

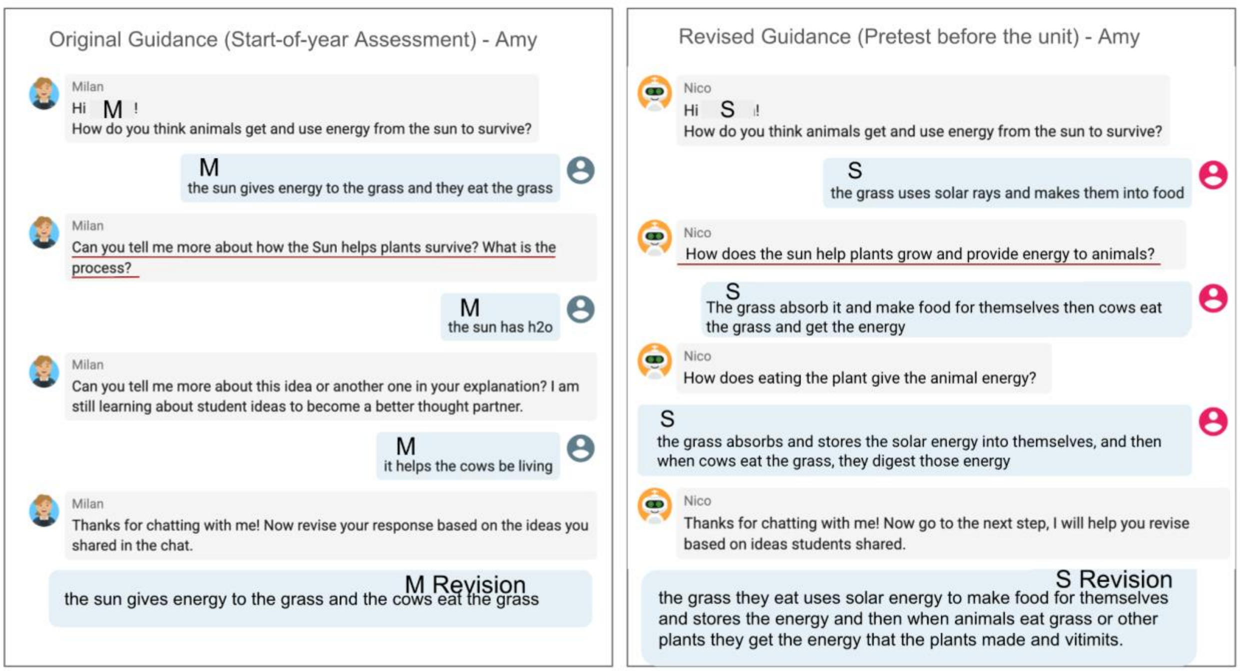

Figure 5. Student dialog examples of Amy's original vs revised guidance.

Student M, receiving the original compound guidance ("Can you tell me more about how the Sun helps plants survive? What is the process?"), struggled to engage with the question's complexity. M's first response showed confusion ("the sun has h2o"), conflating the sun with water. When prompted again with a generic elicitation, M provided only a vague statement ("it helps the cows be living"). M's revised explanation remained at a surface level, merely repeating the initial observation without explaining any mechanism: "the sun gives energy to the grass and the cows eat the grass."

In contrast, Student S, receiving Amy's revised guidance ("How does the sun help plants grow and provide energy to animals?"), immediately engaged with the mechanism. S explained

"The grass absorb it and make food for themselves then cows eat the grass and get the energy," demonstrating understanding of energy transfer. When Amy's revised second prompt asked "How does eating the plant give the animal energy?" S traced the complete energy flow: "the grass absorbs and stores the solar energy into themselves, and then when cows eat the grass, they digest those energy." S's revised explanation integrated these ideas into a coherent mechanistic account: "the grass they eat uses solar energy to make food for themselves and stores the energy and then when animals eat grass or other plants they get the energy that the plants made and vitimits."

The contrast illustrates Amy's revision strategy: the original compound question with abstract process-naming ("tell me more... what is the process?") created a cognitive barrier that prevented M from articulating mechanisms. Amy's revised guidance replaced this with a concrete, focused question emphasizing energy use, which successfully prompted S to explain energy transfer and trace multi-step energy flow through the food chain. This shift from vague elaboration prompts to specific mechanism questions directly supported students in making energy connections visible, exactly Amy's pedagogical goal.

4.2.4 Ben's Student Outcomes

To test whether Ben's accessibility-focused revisions (Table 2) improved student responses, I conducted two GEE analyses parallel to Amy's case.

Analysis 1: Four revised items. Filtering to Ben's four revised items yielded 73 guidance-response observations (38 from Start-of-year Assessment, 35 from Unit Pretest). Given Ben's accessibility goal, I examined two outcomes: (1) new ideas of any type (normative, intuitive, or precursor), and (2) new normative ideas specifically. For new ideas, students added significantly more new ideas when receiving revised guidance at Pretest ($M = 0.54$, $SD = 0.72$) compared to original guidance at Start-of-year Assessment ($M = 0.31$, $SD = 0.58$) ($\beta = 0.56$, $SE = 0.22$, $z = 2.55$, $p = .011$; $IRR = 1.75$, 95% CI [1.14, 2.69]). Students added 75% more ideas with Ben's accessibility-focused revisions, indicating that simplifying linguistic complexity supported increased elaboration and engagement. These ideas were identified using NLP models that detect conceptual content based on semantic meaning, not keyword matching, and the revised guidance did not introduce new scientific vocabulary that students could simply copy.

However, for new normative ideas specifically, the improvement was not statistically significant (Start-of-year: $M = 0.11$, $SD = 0.32$; Pretest: $M = 0.19$, $SD = 0.44$; $\beta = 0.41$, $SE = 0.28$, $z = 1.46$, $p = .144$; $IRR = 1.51$, 95% CI [0.87, 2.61]). While numerically more students added normative ideas with revised guidance, this difference did not reach significance. Dialog round was not a significant predictor for either outcome.

Analysis 2: Revised vs. unchanged comparison. Using all 252 guidance-response observations (63 students × 2 timepoints × 2 rounds), the Timepoint × IsRevised interaction was marginally significant ($\beta = 0.41$, $SE = 0.24$, $z = 1.71$, $p = .087$), with revised items showing somewhat larger improvement ($IRR = 1.63$) compared to unchanged items ($IRR = 1.25$).

The pattern of results (significant increase in idea elaboration but not specifically normative ideas) aligns with Ben's pedagogical goal of accessibility. His linguistic simplifications successfully reduced barriers to participation for his students who are multilingual and just started receiving science instruction, enabling them to engage more fully with the dialog and add more ideas overall. However, accessibility alone did not guarantee acquisition of scientific mechanism, suggesting that supporting emergent bilinguals in science requires both linguistic accessibility (which Ben provided) and additional scaffolding for scientific reasoning with instruction in other parts of the unit (which the dialog alone may not have fully provided).

4.2.5 Ben's Student Cases

Student examples illustrate how Ben's accessibility-focused revisions supported different types of engagement. Figure 6 compares two students' dialog experiences with original versus revised guidance for the same idea (Idea 11a: animals directly use sun's energy).

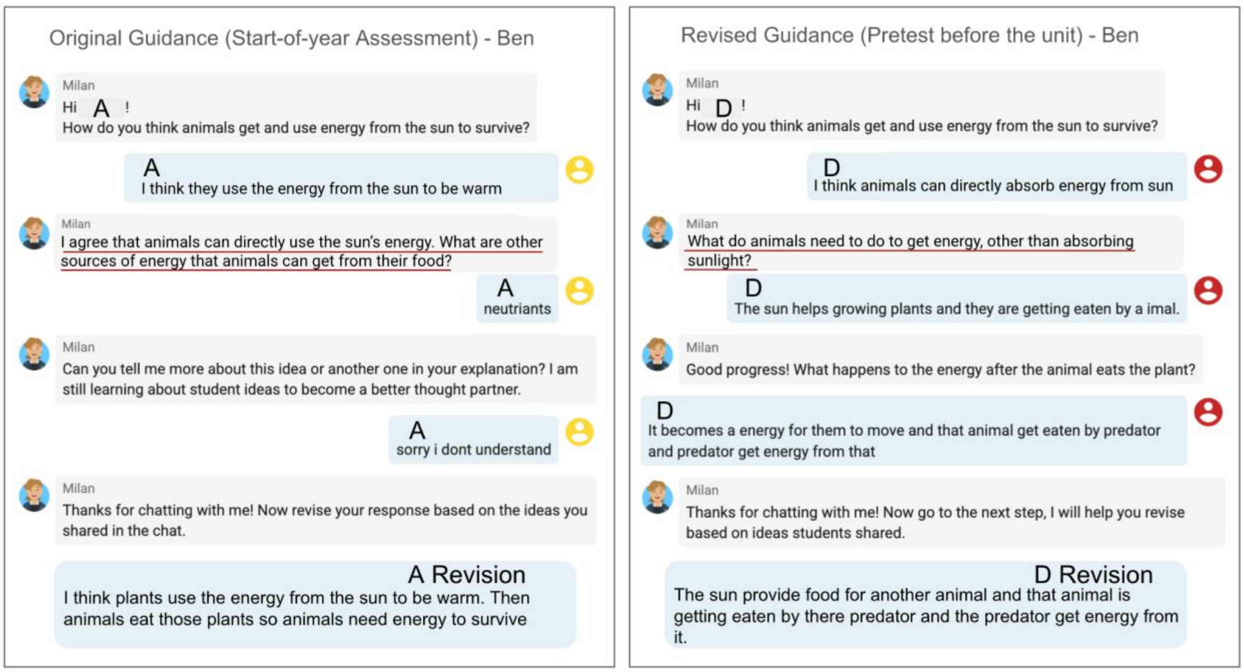

Figure 6. Student dialog examples of Ben's original vs revised guidance.

Student A, receiving the original guidance with complex syntax and abstract language ("I agree that animals can directly use the sun's energy. What are other sources of energy that animals can get from their food?"), provided minimal responses that did not engage with the question's intent. A's first response was a single word ("nutrients"), and when prompted with a generic follow-up, A wrote "sorry i dont understand," explicitly indicating that the language created a comprehension barrier. A's revised explanation showed minimal development: "I think plants use the energy from the sun to be warm. Then animals eat those plants so animals need energy to survive." While A attempted to connect sun, plants, and animals, the explanation remained vague without mechanistic detail.

In contrast, Student D, receiving Ben's revised guidance with simplified syntax and concrete language ("What do animals need to do to get energy, other than absorbing sunlight?"), engaged immediately and productively. D explained “The sun helps growing plants and they are getting eaten by animals,” demonstrating understanding that energy moves from plants to animals. When Ben’s revised distinguishing prompt asked “What happens to the energy after the animal eats the plant?” D traced energy transformation through trophic levels: “It becomes energy for them to move and that animal get eaten by predator and predator get energy from that.” D’s revised explanation demonstrated sophisticated mechanistic understanding: “The sun provide food for another animal and that animal is getting eaten by there predator and the predator get energy from it.”

The contrast illustrates Ben's revision strategy: the original guidance contained multiple embedded clauses, compound sentences, and abstract academic language (“other sources of energy”) that overwhelmed emergent bilingual sixth-graders’ working memory. Ben’s revised guidance simplified to single-clause questions with concrete action verbs (“need to do,” “happens to”) and everyday language. This linguistic simplification removed the comprehension barrier, enabling D to engage with scientific content and demonstrate mechanistic reasoning about energy transfer. Ben's revisions enabled D to demonstrate mechanistic reasoning that A could not access with the original guidance, indicating the barrier was linguistic complexity rather than conceptual inability. Notably, D’s explanation, while showing strong mechanistic understanding, contained grammatical errors typical of emergent bilinguals (“there predator” instead of “their predator”), confirming that the issue was linguistic access, not scientific reasoning capacity.

4.2.6 Summary: Goal-Driven Improvements Across Teachers

Analysis of Amy's and Ben's customization processes reveals a consistent pattern: teachers' pedagogical goals shaped revision strategies and produced student learning outcomes aligned with those goals.

Amy's goal of making energy connections visible led to revisions emphasizing process questions, dynamic flow language, and focused causal prompts. Student outcomes aligned with this goal: more than twice as many normative, mechanistic ideas added (IRR = 2.18, $p < .001$), with revised items showing significantly greater improvement than unchanged items ($p = .013$). Ben's goal of making science accessible for emergent bilinguals led to revisions emphasizing concrete action verbs, single-focus questions, and everyday language. Student outcomes aligned with this goal: 75% more ideas added overall (IRR = 1.75, $p = .011$), indicating that linguistic simplification successfully reduced barriers to participation, though gains in normative ideas specifically did not reach significance ($p = .144$).

Student examples illustrated how these goal-aligned revisions supported different types of engagement: Amy's process-focused questions enabled mechanistic tracing of energy flow (Student S), while Ben's linguistically simplified questions enabled emergent bilinguals to access and articulate scientific content (Student D). These complementary customizations demonstrate how different pedagogical goals produce different but valid learning trajectories, each addressing specific student needs.

This systematic alignment, from stated goals to revision patterns to student outcomes, demonstrates that teachers' pedagogical goals functioned as organizing principles, creating coherence across design decisions and producing improvements in specific, goal-aligned learning outcomes. Both teachers' revisions measurably improved student responses, but in outcomes reflecting their distinct pedagogical priorities and students' specific needs.

4.3 RQ2: Instructional pathways supporting knowledge integration in two cases

To examine how teachers' co-designed classroom activities worked together with AI dialogs to support knowledge integration, I analyzed learning trajectories across four timepoints in two classrooms. Each teacher co-designed activities based on their existing practices aligned with their pedagogical goals: Amy co-designed a drawing activity to make ecosystem connections visible; Ben co-designed scaffolded labs to make mechanistic details accessible. Both pathways supported students in integrating ideas about photosynthesis and energy transfer, though through different instructional emphases.

4.3.1 Amy's Case: Making Connections Visible Through Drawing

Activity co-design

Consistent with Amy's goal of making connections visible, Amy and the first author co-designed an ecosystem drawing activity that required students to visually represent relationships among organisms in their school garden ecosystem. After Lesson 4 (photosynthesis and energy transfer) and the embedded test, students received a template showing their school garden's raised beds. Amy's choice to use the school garden (rather than a generic ecosystem) reflected her second pedagogical goal of making science personally relevant, anchoring the activity in students' lived experiences. Students were then asked to: (1) draw at least three organisms (producers, consumers, decomposers), (2) label each organism's role, and (3) draw arrows showing how energy and matter flow among organisms, labeling each arrow as “gets energy from,” “gets matter from,” or “gets energy and matter from.”

This activity reflected Amy's connection goal. As she explained in a pre-implementation meeting: “I wanted them to have concrete ways to consolidate abstract concepts. The dialog gets them thinking about photosynthesis and energy transfer, but I need them to see it as a system. The arrows force them to think about who's connected to who and why.” The activity design operationalized her pedagogical priority: students could not complete the task without both distinguishing among organism types AND explicitly connecting them through labeled relationships.

Amy's instruction during the drawing activity emphasized two KI processes: distinguishing among ecosystem components and connecting ideas through relationships. When introducing the activity, she explicitly guided students through distinguishing: “First, think about what each organism does. Is it making food from sunlight? Eating other organisms? Breaking things down? That tells you its role.” As students worked, she pressed them to connect: “Don't just draw organisms. Show me how the energy moves. Where did THAT organism get its energy?” Her feedback on student drawings consistently focused on relationships: “You have the organisms, but I don't see how the sun's energy gets to the rabbit. Add the arrows to show the connections."

Quantitative outcomes

Students in Amy's classroom demonstrated significant KI gains both within the dialog (initial to revised explanations) and across the instructional unit (pretest to delayed posttest). Table 4 presents mean KI scores at each timepoint for both initial and revised explanations.

Table 4. Knowledge Integration Scores Across Timepoints (Amy's Case)

Timepoint	Initial Mean KI (SD)	Revised Mean KI (SD)	Gain	N
Pretest	2.35 (0.52)	2.65 (0.48)	+0.30	69
Embedded test	2.58 (0.51)	2.87 (0.49)	+0.29	69
Posttest	2.78 (0.50)	3.07 (0.47)	+0.29	68
Delayed posttest	2.52 (0.61)	2.80 (0.57)	+0.28	64

Note. Initial explanation = before receiving adaptive dialog guidance; Revised explanation = after two rounds of adaptive guidance. Delayed posttest occurred six months after instruction.

A cumulative link mixed model (CLMM) with random intercepts for students revealed two significant main effects. First, the dialog produced consistent KI gains: students scored significantly higher on revised explanations than initial explanations across all four timepoints ($\beta = 1.63$, SE = 0.32, $z = 5.05$, $p < .001$), with within-dialog gains of 0.28-0.30 KI levels (Figure 7). The insignificant Dialog × Timepoint interaction ($p = .15$) indicates the dialog was effective throughout instruction. Second, students showed significant KI gains from pretest to posttest ($\beta = 1.62$, SE = 0.29, $z = 5.53$, $p < .001$), with scores remaining significantly higher at delayed posttest than the pretest six months later ($\beta = 0.58$, SE = 0.30, $z = 1.94$, $p = .05$).

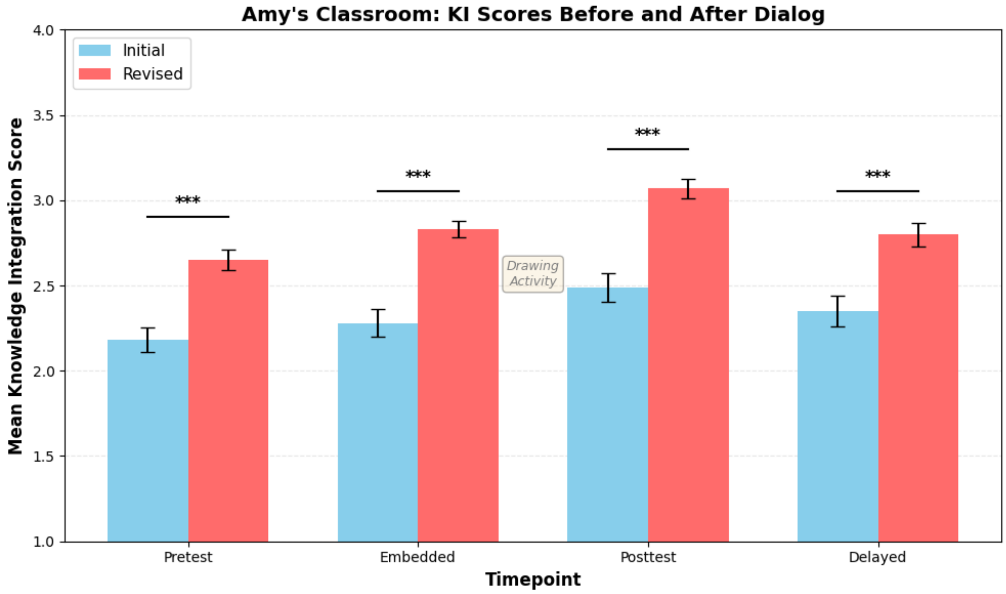

Figure 7. KI scores before and after the dialog (with Revised Guidance) across instruction in Amy's case

Examining how the dialog supported students at different instructional points reveals the synergy between Amy's drawing activity and the dialog. At pretest (before the drawing), the dialog primarily helped students move from KI 2 (isolated ideas) to KI 3 (one link between target ideas). For example, connecting “plants do photosynthesis” with “animals eat plants.” By posttest (after the drawing), the dialog's effectiveness shifted: students more frequently moved from KI 3 to KI 4 (multiple links), with 32% reaching KI 4 compared to 17% at pretest. For instance, Student J entered posttest at KI 3 with “plants make glucose from sunlight and animals eat plants to get energy” and, after dialog prompts, added cellular respiration: “Animals break down the glucose through cellular respiration to get energy for running and growing.” The drawing activity and unit instruction appeared to consolidate students' understanding of ecosystem relationships, enabling the dialog to support fuller integration.

Examining specific ideas confirmed the impact of Amy's connection-focused pathway. Energy transfer, the core normative idea to understand ecosystem energy flow, appeared in 60% of students' revised explanations at pretest, increased to 85% at posttest, and remained at 70% at delayed posttest, demonstrating strong growth and retention (See Figure 8). Statistical significance for added ideas tested using Generalized Estimating Equations (GEE) with binomial distribution. The mechanistic idea of animal cellular respiration showed the largest gains: 37.92% of students added this idea from pretest to posttest ($p < .001$), with the sharpest increase (27.12%, $p < .001$) occurring from embedded to posttest, after the drawing activity that required tracing energy inside organisms. Photosynthesis ideas also grew significantly, with 28.75% adding these ideas by delayed posttest ($p < .001$). These patterns reflect Amy's sustained instructional emphasis on ecosystem connections and energy flow relationships.

Other mechanistic ideas also showed growth. The idea of photosynthesis products (glucose, oxygen) increased from 5.0% at pretest to 23.3% at posttest, maintaining 18.6% at delayed posttest. Photosynthesis reactants (CO_2, H_2O) idea grew from minimal baseline (1.5%) to 10.2% at embedded and 15.2% at posttest. These mechanistic details reflect students' deepening understanding of photosynthesis processes, which is the foundational knowledge of energy and matter flow. The regular garden visits throughout the unit provided ongoing authentic context for applying and consolidating these ideas, supporting both immediate learning and long-term retention across all idea types.

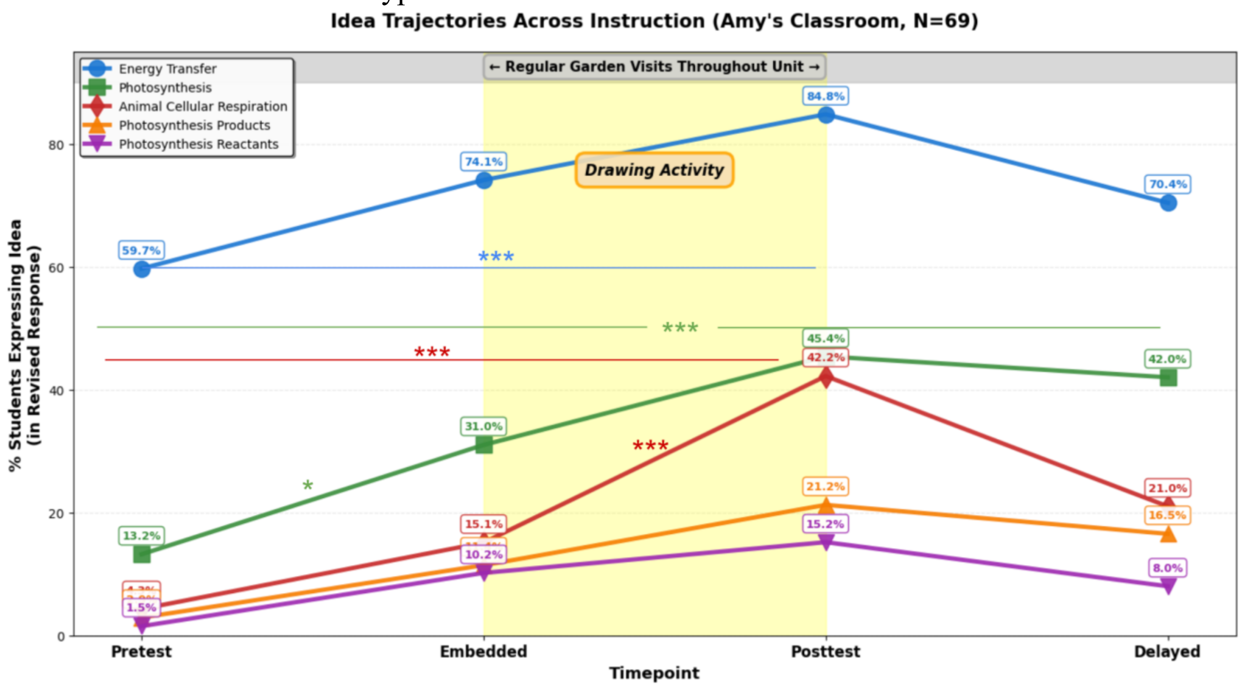

Figure 8. Most added idea across instruction timepoints in the students' revised response after the dialog in Amy's classroom

Note. "Added ideas" = percentage of students expressing an idea at the later timepoint but not the earlier timepoint. Statistical significance for added ideas tested using Generalized Estimating Equations (GEE) with binomial distribution: ****p* < .001, ***p* < .01, **p* < .05.

4.3.2 Ben's Case: Making Science Accessible Through Scaffolded Labs

Activity design

Consistent with his goal of making science accessible, Ben and the first author co-designed what he called "class-based labs" where students collectively observed and discussed simulations on WISE before engaging with the dialog individually. Before Lesson 4 and the embedded test, Ben projected the photosynthesis simulation on the classroom screen and facilitated whole-class observation and discussion. He ran the simulation multiple times, each time asking students to focus on specific aspects: "First time, just watch. What do you see happening? Second time, watch what goes into the plant. What does the plant need? Third time, watch what comes out. What does the plant make?"

This activity reflected Ben's accessibility goals for his 6th graders. As he explained: "A lot of them haven't had any science instruction up till now. They're not the best at teaching themselves things. So I can't just say 'go explore the simulation.' I need to build shared understanding first, what are we looking at? What matters? What's the science here?” The whole-class structure provided scaffolding through collective sense-making before individual application, operationalizing Ben's priority of ensuring complex concepts were comprehensible to students with limited science backgrounds.

Ben's instruction emphasized two KI processes: eliciting students' initial observations and discovering mechanistic details through guided attention. He elicited by asking open-ended observation questions: “What do you notice? What's changing?” Students called out observations (“The red molecule is coming out!” “The stacks are charging!”), and Ben recorded these on the board without initially judging them. Then he guided students to discover mechanisms: “You said the red molecule. What do you think those are? Let's look at our labels [pointing to the simulation key]. It says 'oxygen atom.' Where do they come from?” Through this scaffolded questioning, students collectively discovered that photosynthesis takes in carbon dioxide and water and produces glucose and oxygen.

Ben's approach minimized linguistic demands by using the simulation as a shared visual reference. Instead of asking students to read complex text, he asked them to observe and discuss what they saw: “How many carbon atoms are coming in? Show me with your finger. Now where do they go?” This visual-first approach aligned with his accessibility goal, allowing emergent bilinguals to engage with complex content without language barriers.

Quantitative outcomes

Students in Ben's classroom demonstrated significant KI gains within the dialog (from initial to revised explanations) and across the unit through posttest. However, statistical tests indicated that KI scores at delayed posttest six months later were not significantly different from at pretest. Table 6 presents mean KI scores.

Table 6. KI Scores with the dialog (Initial vs Revised) Across Timepoints (Ben's Case)

Timepoint	Initial explanation *M* (*SD*)	Revised explanation *M* (*SD*)	KI Gain	N
Pretest	2.07 (0.77)	2.32 (0.59)	+0.25	63
Embedded	2.28 (0.69)	2.63 (0.52)	+0.35	63
Posttest	2.40 (0.68)	2.69 (0.50)	+0.29	62
Delayed	2.21 (0.61)	2.48 (0.57)	+0.27	58

Note. Initial explanation = before receiving adaptive dialog guidance; Revised explanation = after two rounds of adaptive guidance. Dialog gain = difference between revised and initial scores. Delayed posttest occurred six months after instruction.

A cumulative link mixed model (CLMM) revealed two significant main effects. First, students scored significantly higher on revised explanations than initial explanations ($\beta = 1.75$, $SE = 0.39$, $z = 4.49$, $p < .001$), with within-dialog gains of 0.25-0.35 KI levels (Figure 9). The insignificant Dialog × Timepoint interaction ($p = .22$) indicates consistent dialog effectiveness. Second, KI scores increased significantly from pretest to embedded test ($\beta = 0.25$, $SE = 0.11$, $z = 2.27$, $p = .023$) and from pretest to posttest ($\beta = 0.35$, $SE = 0.17$, $z = 2.06$, $p = .039$). At delayed posttest six months later, scores remained numerically higher than pretest but this difference was not statistically significant ($\beta = 0.08$, $SE = 0.12$, $z = 0.67$, $p = .50$).

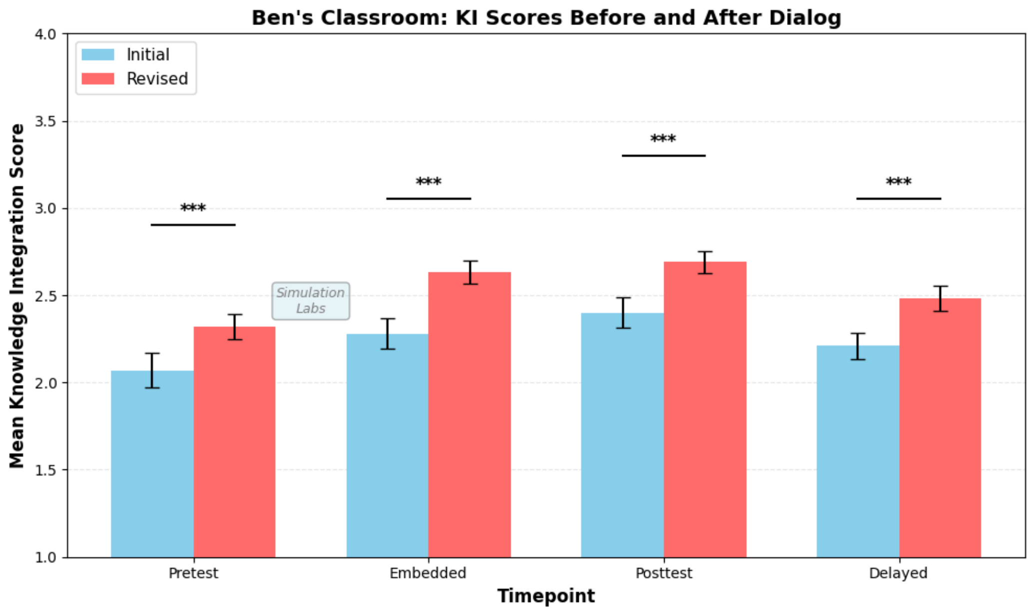

Figure 9. KI scores before and after the dialog (with Revised Guidance) across instruction in Ben's case

Examining how the dialog supported students at different instructional points reveals how Ben's scaffolded labs prepared students for more sophisticated dialog interactions. At pretest (before the labs), the dialog primarily helped students move from KI 1 (off-topic responses like "idk") to KI 2 (vague observations about food chain like "plants use sunlight to make food"). By embedded test (after the labs), more students entered at KI 2 with the knowledge of how sun helps plants survive learned from whole-class simulations, and the dialog's effectiveness shifted toward helping them connect this knowledge: 57% reached KI 3 at embedded compared to 37% at pretest. For example, Student K entered embedded at KI 2 with "plants use CO2 and water to make glucose" and, after dialog prompts, connected to energy transfer: "When animals eat the plants they get that energy from the glucose." The labs built mechanistic foundations that the dialog could then integrate into broader ecosystem understanding.

Examining specific ideas confirmed the impact of Ben's accessibility-focused pathway (Figure 10). The mechanistic idea of photosynthesis reactants (CO_2 and H_2O) showed the largest gain: 28.75% of students added these mechanistic details from pretest to embedded ($p = .032$), directly following Ben's whole-class scaffolding of molecular observations. The mechanistic idea of photosynthesis products (glucose and oxygen) also increased significantly during this period

(21.5%, p = .041). By posttest, 25.50% had added general photosynthesis understanding (p = .038). Energy transfer ideas increased substantially from 36.5% at pretest to 50.2% at posttest, with noticeable growth following the garden visit (40.0% at embedded to 50.2% at posttest, representing 19.15% growth, p = .087). Animal cellular respiration remained relatively low throughout (1.6% at pretest, peaking at 11.7% at posttest), reflecting that Ben's sixth-graders were still developing foundational photosynthesis concepts and had limited opportunities to explore animal energy processes. These patterns reflect Ben's instructional emphasis on making mechanistic details visible and comprehensible through scaffolded observation, with the single observational garden visit providing concrete context and the dialog helping students connect lab-learned details to broader energy transfer concepts.

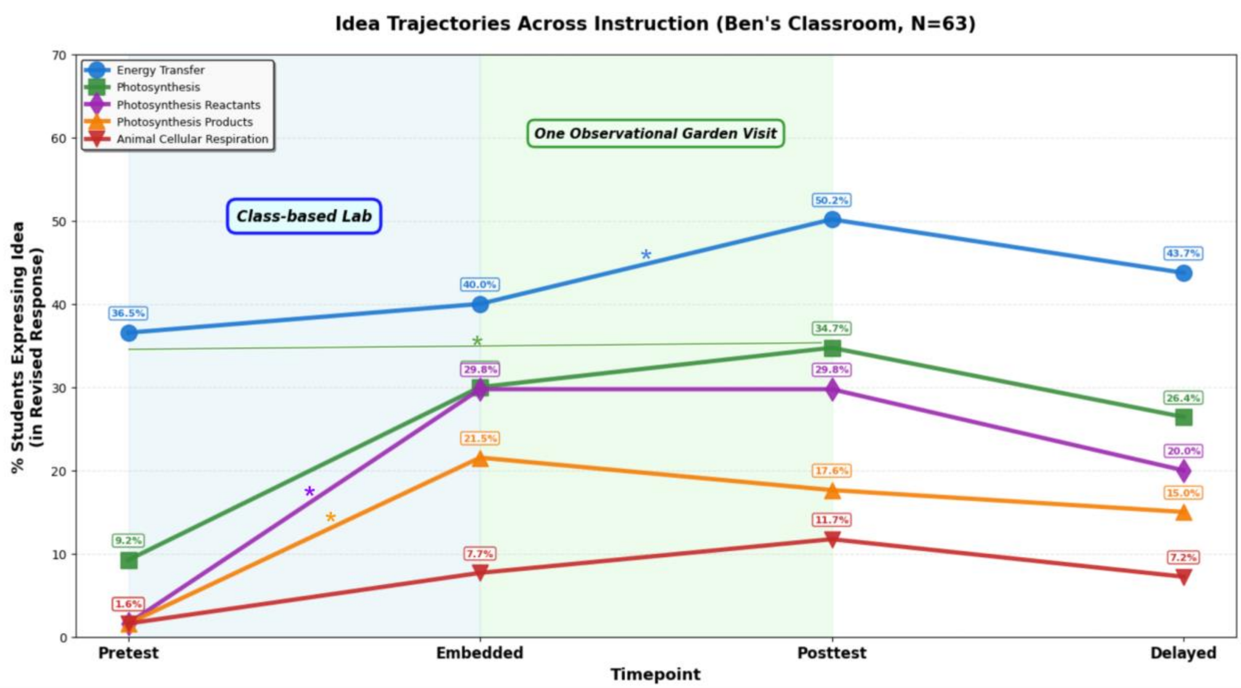

Figure 10. Most added idea across instruction timepoints in the students' revised response after the dialog in Ben's classroom

Note. "Added ideas" = percentage of students expressing an idea at the later timepoint but not the earlier timepoint. Statistical significance for added ideas tested using Generalized Estimating Equations (GEE) with binomial distribution: ****p* < .001, ***p* < .01, **p* < .05.

4.3.3 Summary: Activity-Dialog Synergies Supporting Knowledge Integration

Both teachers created instructional pathways supporting significant knowledge integration gains, with different temporal dynamics reflecting their pedagogical goals and emphases.

Amy's connection-focused pathway produced sustained retention: KI gains remained significant at delayed posttest ($\beta = 0.58$, $p = .05$), and the energy transfer idea appeared in 70% of students' revised explanations six months after instruction. Energy transfer was also consistently retained across timepoints (14-23% of students). Animal cellular respiration idea showed large gains (37.92% added from pretest to posttest, p < .001), particularly after the drawing activity (27.12% added from embedded to posttest, $p < .001$). Students also developed mechanistic understanding of photosynthesis details, with photosynthesis products appearing in 23% of posttest responses and photosynthesis reactants in 15%, indicating that the drawing activity and regular garden visits supported integration of both ecosystem-level relationships and molecular-level mechanisms. In post-implementation interviews, Amy attributed these sustained gains to the repeated opportunities for students to distinguish and connect ideas: “I think the drawing really helped them see the relationships...they weren't just listing facts anymore, they were actually showing how energy moves through the system.” Amy observed that students who initially struggled with fragmented knowledge showed marked improvement: “I saw students connecting what they learned back to the previous ecology lesson we did earlier this year.”

Ben's accessibility-focused pathway produced strong initial gains in mechanistic understanding: photosynthesis reactants increased 28.75% after scaffolded labs ($p = .032$), and products increased 20.03% ($p = .041$). Animal cellular respiration remained relatively low (peaking at 12% at posttest), reflecting that sixth-graders were still building foundational photosynthesis concepts. Overall KI gains were not sustained at delayed posttest ($\beta = 0.08$, $p = .50$), and fewer students maintained integrated explanations six months later. In post-implementation reflections, Ben recognized both the success and limitations of his approach: “The whole-class labs really worked. They got the CO2 and water, they got the sugar and oxygen. But I wonder if they needed more time afterward to connect it all together.” He observed that his sixth graders, many encountering formal science instruction for the first time, successfully acquired foundational mechanistic knowledge but noted: “Some of them could tell me what goes in and what comes out, but connecting that to the bigger energy transfer story...that's still developing.”

Each case demonstrates how a teacher's pedagogical goal organized design decisions across multiple levels. Amy prioritized integrated ecosystem understanding through activities emphasizing relationships; outcomes reflected durable ecosystem-level thinking. Ben prioritized accessible entry for sixth-grade multilingual learners encountering formal science instruction for the first time; his scaffolding supported comprehension of mechanistic photosynthesis details, creating foundational knowledge some students continued building upon. The pattern suggests that instructional pathways emphasizing different KI processes produce different learning trajectories: pathways emphasizing distinguishing and connecting (Amy's approach) may inherently support retention by building integrated mental models, while pathways emphasizing eliciting and discovering (Ben's approach) effectively establish initial comprehension but may require additional consolidation activities for long-term integration.

In both cases, activities and dialogs functioned synergistically. Amy's drawing activity prepared students to move from single connections (KI 3) to multiple linked ideas (KI 4) when the dialog prompted elaboration, while regular garden visits throughout the unit provided ongoing authentic context for consolidating ecosystem concepts; Ben's scaffolded labs built mechanistic foundations that enabled students to connect photosynthesis details to how sun helps plants survive (KI 2 to KI 3) when the dialog asked about mechanisms, and the single observational garden visit offered concrete grounding for energy transfer ideas. Pedagogical goals created coherence across this design: goals shaped how teachers revised dialog guidance, what activities they co-designed, which KI processes they emphasized during instruction, and which learning patterns emerged. When goals consistently organized design decisions across multiple levels, significant learning resulted, though specific outcomes reflected teachers' distinct priorities.

5. Discussion

This study examined how middle school science teachers' pedagogical goals shaped their customization and enactment of an AI adaptive dialog embedded in a photosynthesis and ecosystems unit. The central finding is that teachers developed distinct pedagogical goals rooted in their noticing of their students' ideas and needs, and that these goals, shaped by their engagement with KI pedagogy through the RPP, created coherence across dialog guidance refinements and classroom activity design. Although the two teachers pursued different goals, both produced instructional customizations that yielded significant student knowledge integration gains, revealing the richness of KI-informed pedagogy across diverse goal preferences (Table 7).

Table 7. Dialog Effectiveness and Customization Impact Across Two Cases

		Amy	**Ben**
Student context	Grade/demographics	7-8th grade, 80% multilingual	6th grade, 85% multilingual, limited prior science
	Pedagogical goal	Making connections visible	Making science accessible
Dialog effect (all timepoints)	Within-dialog KI gain	$\beta = 1.63$, $p < .001$	$\beta = 1.75$, $p < .001$
	Dialog × Timepoint interaction	Not significant (Dialog effect stable across all 4 timepoints)	Not significant (Dialog effect stable across all 4 timepoints)
Customization effect (RQ1)	Revised vs. original guidance (Analysis 1)	IRR = 2.18, $p < .001$ (2× more normative ideas)	IRR = 1.75, $p = .011$ (75% more new ideas overall)
	Revised vs. unchanged guidance (Analysis 2)	Significant differential ($p = .013$)	Marginal differential ($p = .087$)

Goal-aligned customizations (RQ2)	Primary mechanistic idea gain	Animal cellular respiration +38% from pretest to posttest ($p < .001$)	Photosynthesis +26% from pretest to posttest ($p = .038$)
	Activity-dialog synergy	Animal cellular respiration +27%*** immediately after drawing activity	Photosynthesis reactants (+29%*)/Products (+20%*) spiked immediately after class-based labs
	Instructional KI gains (pretest to posttest)	$\beta = 1.62$, $p < .001$	$\beta = 0.35$, $p = .039$
	KI gain at delayed posttest	Significant ($\beta = 0.58$, $p = .05$)	Not significant ($\beta = 0.08$, $p = .50$)
Interpretation	Dialog contribution	Robust gains across diverse contexts	Robust gains across diverse contexts
	Customization contribution	Directed learning toward ecosystem-level integration (cellular respiration, energy transfer connections) while amplifying mechanistic reasoning	Made foundational photosynthesis concepts accessible (reactants, products) by removing linguistic barriers

Teachers' pedagogical goals were grounded in their noticing of specific patterns in students' ideas and learning needs (Luna, 2018; Wilkerson et al., 2016). Amy noticed that her students held disconnected ecosystem ideas: they could name the parts but could not trace energy moving through the system. She diagnosed a need for opportunities to make relationships among ideas visible. Ben noticed that his emergent bilingual students were shutting down when encountering complex sentence structures in the dialog, and diagnosed both a linguistic accessibility problem and a scaffolding problem for students encountering formal science for the first time. Critically, both teachers drew on KI design principles developed through RPP workshops and courses to translate these observations into actionable goals: Amy's noticing mapped onto making thinking visible, Ben's onto making science accessible. This connection between classroom noticing and KI principles gave each teacher a coherent organizing framework they could carry across multiple levels of instructional decision-making.

Analyzing customizations across dialog guidance, activity design, and classroom enactment revealed that each level of customization aligned with and reinforced the same pedagogical goal (Table 8). Amy's customizations consistently emphasized making connections visible: she shifted guidance from categorization to process questions requiring students to trace energy movement, designed a drawing activity requiring relationship arrows connecting organisms, and pressed students during instruction to "show me how the energy moves." Ben's customizations consistently emphasized making science accessible: he simplified syntactic complexity in dialog guidance, designed whole-class simulation labs building shared understanding before individual work, and asked students "what do you notice?" and "where do molecules come from?" during instruction. Amy reflected that the drawing "really helped them see the relationships...they weren't just listing facts anymore, they were actually showing how energy moves through the system," while Ben noted that "the whole-class labs really worked...but I wonder if they needed more time afterward to connect it all together". These observations confirmed goal-aligned customization shaped both what students learned and where further development remained.

Table 8. How Pedagogical Goals Created Coherence Across Customizations

Teacher	Noticing of Student Ideas	Diagnosing student needs	Main Pedagogical Goal (by KI lens)	Guidance Customizations (Example)	Activity Customizations	Instructional Emphasis
Amy	Students listed disconnected and abstract concepts in ecosystems without tracing relationships	Students need opportunities to make connections among ideas visible	Making connections visible	Revised toward asking mechanisms: "What type of energy?" → "How does eating the plant give the animal energy?"	Ecosystem drawing requiring relationship arrows connecting organisms	Emphasized **distinguishing** organism roles and **connecting** through energy flow
Ben	Emergent bilinguals shut down with complex sentence; limited prior science knowledge	Students need linguistic scaffolding and shared discovery before independent work	Making science accessible through scaffolded discovery	Made guidance more accessible: "I agree that animals can directly use the sun's energy. What are other sources of energy..." → "What do animals need to do to get energy, other than	Whole-class simulation labs with scaffolded collective observation before individual work	Emphasized **eliciting** observations and **discovering** mechanisms through guided whole class discussion

				absorbing sunlight?"		

The two teachers' different pedagogical goals reflect different KI design principles, and these different principles led teachers to emphasize different KI processes in their customizations. Amy's connection goal led her to emphasize distinguishing and connecting: process questions tracing energy movement, drawing activities requiring relationship arrows, and instructional moves pressing students to "show me how the energy moves." Ben's accessibility goal led him to emphasize eliciting and discovering: simplified guidance reducing cognitive load, whole-class simulation labs building shared observations, and instructional moves asking "what do you notice?" and "where do molecules come from?" These different process emphases produced different but equally valid learning outcomes. Amy's students showed significant KI gains through delayed posttest ($\beta = 0.58$, $p = .05$), with 70% expressing energy transfer ideas and 38% adding cellular respiration six months after instruction, demonstrating ecosystem-level integration. Ben's accessibility-focused customizations increased new ideas by 75% (IRR = 1.75, $p = .011$), successfully removing participation barriers and enabling emergent multilingual students to engage with complex content many had never encountered. His students developed foundational mechanistic knowledge (reactants, products) and, critically, developed approaches to engaging with science topics they could apply to future learning. For sixth-graders in their first formal science instruction, this accessibility-focused customization prevented exclusion from complex content while building both conceptual foundations and productive learning strategies.

This customization coherence extends prior research on teachers adapting curriculum materials (Remillard, 2000; Fischer et al., 2013; Matuk, McElhaney, et al., 2016; Severance et al., 2016) and RPP-supported teacher customization (Gerard et al., 2011; Billings & Linn, 2024). This study enriches our understanding by illustrating how goal-driven customization produces learning gains. These cases show that when pedagogical goals are organized by KI design principles, the same goal can structure customizations across curriculum customization, classroom practices, and interpretation of student progress simultaneously. This coherence distinguishes productive customization from isolated adaptation. Goal-driven coherence thus functions as a flexible rather than prescriptive mechanism: when teachers' goals are responsive to their students' ideas and structured by a shared pedagogical framework, they lead to significant knowledge integration results across diverse classroom contexts (Linn et al., 2004; Penuel et al., 2011).

6 Implications

Both cases demonstrated that teachers contribute essential expertise about their students and contexts to the RPP. Teachers identified new guidance problems, designed locally relevant activities, and enacted instruction that amplified the value of the AI system. AI systems should therefore be built for teacher adaptation rather than standardized implementation: providing teachers with actionable student response data (Wiley, Dimitriadis & Linn, 2024), editable guidance templates, and platform flexibility for locally designed activities. In this RPP, researchers prepared annotated student dialog samples showing which ideas were detected and which prompts needed revision, allowing teachers to see patterns across their classrooms and act on them. Automating these resources within the AI tools themselves is a productive direction for future development. Critically, effective customization did not require teachers to adopt entirely new approaches: Amy had used drawings before, Ben had simplified language before. This RPP provided twice-yearly workshops, technical support, and multiple opportunities to develop a deep understanding of KI pedagogy and WISE platform, and to help each teacher formalize these practices, coordinate them with the AI dialog, and make their pedagogical logic explicit. Professional learning that helps teachers recognize how their established practices connect to new technologies, rather than training predetermined implementation, is likely to produce more sustained and contextually responsive customization (Gerard et al., 2011; Luna, 2018; Lowell, 2024).

A broader implication concerns how AI and teacher expertise interact in classroom learning environments. The AI dialog and teacher-designed practices served complementary rather than redundant functions: the dialog elicited students' existing ideas at scale and provided personalized adaptive guidance; teacher activities consolidated and connected those dialog-elicited ideas through locally meaningful sense-making grounded in knowledge of specific students and communities (Barton & Tan, 2009). Amy's drawings organized dialog-surfaced ideas into visible relational systems; Ben's labs grounded abstract dialog concepts in shared embodied observation. Neither component alone produced the observed gains; effectiveness emerged from their coherent integration (Tansomboon et al., 2017; Penuel et al., 2023). This raises a critical design question: How can AI support teachers in doing what only teachers can do? The answer involves designing systems that enable teachers to create locally responsive practices grounded in knowledge of their students and communities.

7 Limitations and Future Directions

Two design features limit causal inference. For RQ1, timepoint and guidance version are confounded: all students received original guidance in September and revised guidance in December, with three months of intervening science instruction (Amy: water and life science; Ben: cells and cell function). For RQ2, we cannot isolate independent effects of dialog versus classroom activities, as students experienced both as integrated instruction. In both cases, convergent evidence supports inferences about the contributions of customization, but readers should interpret effect sizes with these confounds in mind. Future research using within-classroom randomization or component-isolation designs would enable stronger causal claims.

Generalizability is limited in several ways. Both teachers had dedicated customization time and sustained researcher support unavailable to most teachers, taught multilingual learners in urban public schools, and worked with a single science topic. These factors may have shaped both the goals teachers developed and the forms customization took. Whether goal-driven coherence patterns emerge across other topics, student populations, teacher experience levels, and less intensive professional learning structures remains an open question.

The five-level KI scoring rubric, while validated for measuring integrated understanding, may underrepresent incremental conceptual growth. Ben's students demonstrated significant gains in mechanistic details yet some remained at KI level 2 because they had not yet linked those details into complete ecosystem explanations. The rubric appropriately rewards integration over isolated idea accumulation, but this means foundational learning, particularly important for novice learners, may be underrepresented in KI score gains. Complementary measures such as idea frequency trajectories (Figures 9 and 11) better characterize learning for students building foundational knowledge, and future research could develop intermediate scoring criteria to capture this progression.

Future research should address several questions. First, regarding generalizability across contexts: do goal-driven coherence patterns hold when teachers customize dialogs for other science topics, student populations, and professional learning structures less intensive than a sustained RPP? Second, regarding KI process sequencing: For students at different developmental stages, what is the optimal emphasis and sequencing of the four KI processes (elicit, discover, distinguish, connect)? Finally, this study traced organizing goals that shaped customizations across multiple design levels, but teachers also articulated contextual goals shaping motivation and personal relevance. How these two types of goals interact, and whether their combination produces more durable outcomes than either alone, is a productive direction for future research.

Chapter 4: Designing AI Dialogs for Food Justice

Chapter summary

When students explain why some neighborhoods lack access to fresh food, they draw on community observations, personal experience, and cultural knowledge. When a student writes, "my neighborhood only has food trucks but my friend's neighborhood has a El Rancho Supermercado" they are explaining an aspect of food justice. Interpreting this explanation requires understanding the neighborhood and the difference between food trucks and supermarkets. This chapter discusses how an RPP built an idea rubric and AI Idea Detection model for a food justice Knowledge Integration explanation item. This involves interpreting student ideas, determining which ideas are relevant, designing a AI idea detection model, and creating an AI dialog. The development of the rubric and the dialog is informed by the Knowledge Integration framework (KI; Linn & Eylon, 2011) and Justice-Centered Science Pedagogy (JCSP; Morales-Doyle, 2017).

This chapter reports on the dialog design and implementation. In phase 1, the RPP developed an idea rubric for the food justice item, built the AI idea detection model, and designed guidance for a food justice dialog. The RPP refined the dialog in two iterations. Eight partner teachers representing diverse school contexts contributed to the design, ensuring the rubric and guidance captured the range of ideas students expressed across classrooms. In phase 2, two of the eight teachers implemented the dialog, aligning it with their pedagogical goals. The teachers' customizations are analyzed using KI and JCSP design principles.

Unlike the development of the photosynthesis dialog (Chapter 2), that refined an existing rubric, the food justice dialog required building a rubric from scratch. The food justice rubric captures student reasoning that integrates lived experience with a critique of systemic inequity.

1 Introduction

When students examine food access, they bring scientific knowledge about which vegetables are more drought tolerant, and community knowledge about which neighborhoods near their school have fresh produce and which do not. They have observed which families struggle to afford healthy food. Some can connect those observations to systemic inequity, recognizing that the food insecurity in their neighborhood reflects historical redlining policies that systematically divested resources from communities of color. Some have language for what they witness; many do not. Supporting students to build these connections from personal observation to scientific mechanism to reasoning about the historical and current policies and practices that produced unequal food access is the central challenge of teaching food justice as a social justice science issue (SJSI; Morales-Doyle, 2017). Throughout this chapter, ideas that require students to name and explain these historical and systemic injustice are referred to as structural ideas, as distinct from community-based evidence ideas, which reflect what students can directly observe and collect data about food access in their own neighborhoods. Teaching SJSI not only makes science more relevant (Linn & Hsi, 2000) but also develops students' capacity to use scientific reasoning to understand and address societal challenges (Bradford, Gerard, et al., 2023).

Research-practice partnerships (RPP) have developed promising approaches for addressing this challenge across several SJSI contexts. Gerard, Bradford, Debarger, et al. (2022) demonstrated how RPP collaboration with teachers can produce anti-racism justice curricula grounded in teacher expertise and local classroom contexts. Bradford et al. (2023) extended this work by designing a global climate change and urban heat island curriculum with teachers using an integrated KI and JCSP framework, showing that rigorous disciplinary learning and the development of students' critical consciousness can be pursued together. Building on this curriculum, Bradford et al. (2024) developed NLP-based automatic content scoring models that detect the degree to which students integrate science and social justice ideas in climate change contexts, demonstrating that transformer-based models can produce educationally useful scores for SJSI reasoning across different environmental science contexts. Li, Liao, et al. (2024) showed that teacher-informed idea model expansion can improve NLP detection of student reasoning in SJSI contexts. Taken together, this body of work demonstrates that RPPs can build instructional tools and curricula that are grounded in both pedagogical theory and authentic classroom knowledge.

Yet significant challenges remain, particularly for food access and community garden contexts. Ibourk et al. (2024) document how community gardens serve as powerful sites for ecological caring and justice-oriented science learning, and how deeply students engage with food-related science when it connects to their own communities and lived experiences. However, no prior RPP work has built an AI dialog specifically for food justice reasoning, and the detection challenge for this domain is distinct from prior SJSI work, like the global climate change and urban heat island (Bradford et al., 2023). When students reason about food access, their ideas integrate lived experience, plant photosynthesis and ecosystems, community observation, and reasoning about systemic inequity in ways that do not follow recognizable scientific scripts and that biased training data can easily misinterpret or overlook (Sap et al., 2019; Blodgett et al., 2020; Akgun & Greenhow, 2022). Building a model that can reliably detect this full range of ideas, from a student noticing that their neighborhood only has food trucks to a student naming redlining as the cause, requires first building a shared understanding of what food justice reasoning looks like in middle schoolers' words.

This chapter addresses that challenge in two design-based research phases. In Phase 1, the RPP built a shared understanding of food justice reasoning and developed a reliable idea detection model and AI adaptive dialog (Figure 1) across two iterative design cycles, producing a 22-idea rubric and an NLP model that detects the full range of food justice ideas students express. In Phase 2, two focal teachers with different pedagogical orientations toward food justice brought that dialog into their classrooms, and the RPP examined what each teacher's orientation made possible. Together the two phases address the following research questions:

RQ1: How did the RPP build a shared understanding of food justice reasoning and develop the food justice dialog across two iterative design cycles?

RQ2: How do teachers' pedagogical orientations shape what the food justice dialog can accomplish in their classrooms?

To address these questions, the chapter draws on two development iterations involving 798 and 869 middle school students respectively, qualitative data from pre-implementation meetings, classroom observations, and post-implementation interviews with two focal teachers, and quantitative analysis of student dialog data from Ben's 6th-grade and Amy's 8th-grade classrooms.

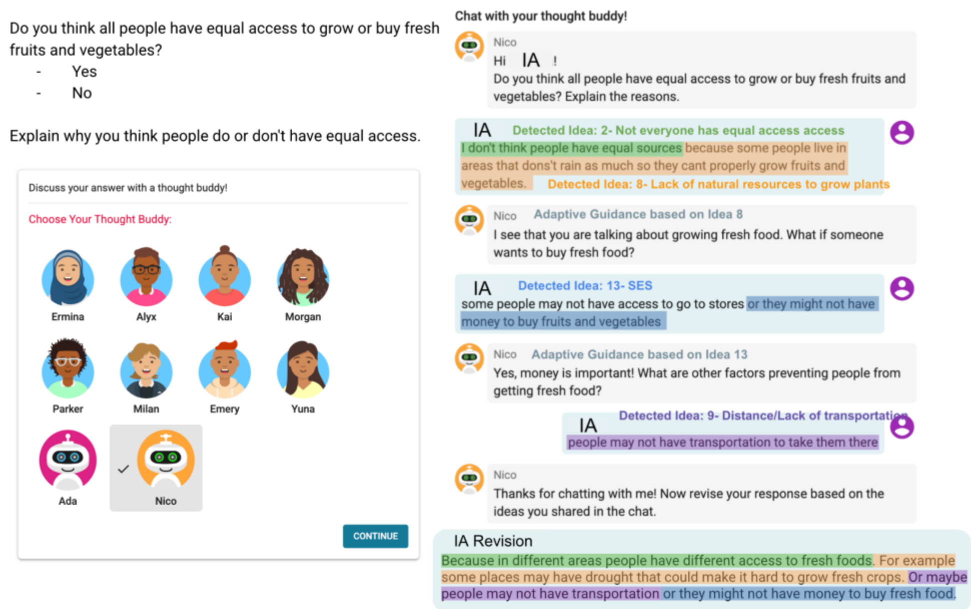

Figure 1. The Food Justice AI Dialog

2 Theoretical Framework

The Knowledge Integration (KI) framework is grounded in the sociocultural view that students' diverse ideas serve as a productive resource for learning rather than misconceptions to be replaced (Smith et al., 1994). KI views learning as a process of sorting out their own ideas alongside new ones they encounter in science class, strengthening their reasoning strategies and increasing the coherence of their understanding over time (Songer & Linn, 1991). This process is guided by four instructional processes: eliciting student explanations, encouraging students to discover additional ideas, guiding students to use evidence to distinguish among their ideas, and asking students to reformulate links among their ideas promotes robust understanding of science concepts (Linn et al., 2023).

To support this process in practice, KI research surfaced four design principles that translate the framework's commitments into instructional decisions (Kali, 2006). Making science accessible to each learner involves connecting scientific concepts to the experiences and contexts that are meaningful to students, rather than relying on generic examples (Linn & Hsi, 2000). Making student thinking visible involves designing activities that require students to articulate and link their ideas with evidence, so that reasoning becomes available for reflection and refinement (Ritchhart & Perkins, 2008). Helping students learn from others involves structuring collaboration so that the diversity of student ideas becomes a resource, not a problem to resolve (Matuk & Linn, 2023). Promoting autonomy involves guiding students' reasoning rather than supplying answers, supporting them in developing confidence in their own reasoning strategies and taking ownership of their learning (Davis, 2003). In this study, these four principles provided the analytical lens for examining teachers' instructional customizations, as described further below.

Drawing on critical pedagogy (Freire, 1970) and culturally relevant pedagogy (Ladson-Billings, 1995), Justice-centered science pedagogy (JCSP; Morales-Doyle, 2017) aims for rigorous disciplinary learning and the development of students' critical consciousness by positioning youth as transformative intellectuals who use scientific practices (e.g., modelling, data analysis, argumentation) to analyze and act upon injustices such as environmental racism. This framework argues that rigorous learning is catalyzed when students work on problems that matter in and to their communities. In this chapter, structural ideas refer to students' reasoning about the historical and current policies, systemic decisions, and racialized practices that produced unequal food access, such as redlining, food apartheid, and corporate concentration of fast food in underserved neighborhoods. Community-based evidence ideas reflect what students can directly observe about food access in their own neighborhoods, such as distance to grocery stores, grocery store availability, and socioeconomic conditions.

Within JCSP, Morales-Doyle (2018) distinguishes among multiple approaches to making science relevant to students' lives, ranging from injected relevance, which attaches real-world examples to a traditional content sequence, to a problem-posing justice-centered approach, which organizes curriculum around SJSI and positions students as agents investigating how science, power, and history intersect. In this study, the two focal teachers' orientations reflected different points along this continuum, shaping which ideas they emphasized and how they structured student engagement with the food justice unit.

KI and JCSP share a foundational commitment rooted in sociocultural theories of learning: both begin from the premise that students arrive in classrooms with valuable prior knowledge. KI recognizes that learners hold varied ideas shaped by observation and lived experience (Matuk & Linn, 2023), and that supporting robust understanding means eliciting those ideas, providing opportunities to discover new ones, using evidence to distinguish among them, and guiding students to form connections (Linn & Eylon, 2011). JCSP extends this commitment to community and cultural knowledge which include the observations students make about their own neighborhoods, the inequities they witness, and the structural and political dimensions of the science issues they study (Morales-Doyle, 2017). Where KI provides the instructional architecture, design principles, and instruction processes that make the vision enactable in classrooms, JCSP provides the conceptual vocabulary to name ideas that KI alone does not yet fully capture: critical consciousness, emotional proximity to the subject of study, and students' positioning as agents who use science to act on their communities.

Taken together, KI and JCSP call for teaching social justice science issues (SJSI; Morales-Doyle, 2017) to elicit and value the full range of their students' ideas, to provide opportunities to discover evidence in a variety of relevant and meaningful contexts, to help students use science to distinguish societal inequities, and to support students in explaining ways to bring about a more just world (Bradford, Gerard, et al., 2023). I combined KI and JCSP to design the food justice AI dialog and curriculum because in combination they surface the diverse ideas students bring to food justice reasoning, provide opportunities to discover community-grounded evidence, and help students use evidence to distinguish the structural factors that produce unequal food access. KI served both as a theoretical and an analytical framework in this study. The four KI design principles, extended by JCSP for the food justice context (Table 1), provided the analytical lens for interpreting teachers' specific instructional moves and customizations.

Table 1. KI Design Principles and Social Justice Science Issues for Food Justice Instruction

KI Meta Principle	KI Design Principle	Social Justice Focus	KI + JCSP for Food Justice
Make science accessible	Connect scientific concepts to relevant and meaningful contexts to motivate	Draw on students' own neighborhood experiences, identities, and cultural	Make food justice accessible by connecting science to students' community

	students to deepen their reasoning (Linn & Hsi, 2000; Linn et al., 2004)	knowledge. Identifies local and personally-relevant examples, not generalized examples (Bang & Medin, 2010)	observations and personal experiences of food access in their own neighborhoods
Make thinking visible	Design explanation activities that require students to articulate and link ideas with evidence, surfacing reasoning that would otherwise remain implicit (Ryoo & Linn, 2012)	Creates space for students to express ideas shaped by their feelings about the issues based on observations of inequity in their own communities (Bradford, Gerard, et al., 2023)	Make food justice reasoning visible by designing activities that require students to articulate both scientific mechanisms and structural observations, including ideas shaped by lived experience
Help students learn from others	Structure collaboration so the diversity of student ideas becomes a resource for refining reasoning (Billings et al., 2024)	Establish norms for safe peer dialogue about racialized and political dimensions of science issues; treats heterogeneity of community knowledge as a shared resource (Morales-Doyle, 2017)	Leverage the diversity of students' community knowledge about food access as a collective resource; building norms for peer dialogue about structural food inequity
Promoting autonomy	Guide students to critique their own and others' ideas, developing investigatory skills (Gerard & Linn, 2022)	Positions students as transformative intellectuals who use science to analyze and act upon injustice in their communities (Morales-Doyle, 2017)	Fostering students' reflections on their observations, perspectives, and potential biases; develop disciplinary understanding to make sense of and address local food inequity

The food justice AI dialog, curriculum, and teacher customizations were designed within RPP to embody this integrated framework. The dialog surfaces ideas students already hold, both scientific and community-based, and keeps pushing them to discover more, connect evidence, and reason toward explanation. The idea rubric was built through iterative RPP design to recognize the full range of ideas this integrated framework predicts students might have, from personal observations about neighborhood food access to structural discrimination policy. The dialog design process documented in RQ1 was an act of collective theorizing about what food justice reasoning looks like in middle schoolers' words, and about which ideas matter enough to detect and respond to. Teacher-designed curriculum activities enacted the extended KI+JCSP design principles, creating the instructional conditions for students to encounter community-grounded evidence, engage in historical and current policies causing local food inequity, and position themselves as agents of change.

3 Research-Practice Partnership Context and Design

This chapter is situated within a long-standing research-practice partnership bringing together middle school science teachers, educational researchers, and computer scientists to develop AI-supported science curricula grounded in teacher expertise and local classroom contexts. The partnership has collaborated across multiple SJSI topics, building shared frameworks for detecting and responding to student reasoning about justice-centered science issues. The food justice work reported here represents the partnership's first effort to build an AI dialog and idea detection model for food justice reasoning specifically, extending the design framework developed for traditional science topics (Chapter 2) and the SJSI dialog work on climate change and urban heat island contexts (Bradford et al., 2024) to a domain where students' ideas integrate lived experience, plant science, community observation, and structural critique.

3.1 Partner Teachers

Eight middle school science teachers participated in designing the food justice dialog and unit over two years (2023-2025), representing diverse school contexts across the western United States. All eight teachers taught the School Garden and Food Justice WISE unit with embedded AI dialogs and contributed to rubric and guidance revision through workshops. Four leading teachers, including Ben and Amy, took on deeper design roles: contributing to 1:1 meetings with the lead researcher during rubric development and adaptive guidance design, piloting materials, and using model accuracy and frequency output to adjust instruction.

Two teachers, Ben and Amy, serve as focal cases in Phase 2 of this chapter, representing the RPP cohort most deeply involved in design, implementation, and customization (Table 2). Their sustained involvement in the design process positions them to reveal what different pedagogical orientations can accomplish with the AI dialog, because they understand both the rubric's intent and the dialog's design logic from the inside. During this study, Ben was pursuing a master's degree in education in which he studied social justice pedagogy frameworks, including Morales-Doyle's (2017) JCSP work, providing him with explicit theoretical grounding for his instructional decisions. Amy is a long-standing RPP member with deep personal investment in social justice teaching and prior experience with the KI framework through years of RPP collaboration.

Table 2. *Partner Teachers: Context and Roles*

Teacher	Grade	School Demographics	RPP Design Role
Amy	7th–8th	93% Hispanic, 3% AAPI; 86% Free/Reduced Lunch	Dialog rubric & guidance design, unit design and customization, piloted materials, focal case (RQ2, RQ1)
Ben	6th	90% Hispanic, 4% Black; 95% Free/Reduced Lunch	Dialog rubric & guidance design, unit design and customization, piloted materials, focal case (RQ2, RQ1)
6 additional teachers	6th–9th	Diverse urban schools	Ran benchmark assessments; piloted unit and dialog; gave feedback during workshops (if attend) (RQ1)

3.2 RPP Design Framework

This study employed design-based research methods (Sandoval & Bell, 2004) to iteratively develop and refine the AI dialog through systematic collaboration in RPP. The RPP design framework originated in a photosynthesis AI dialog co-developed with the same RPP (Chapter 2). Figure 2 illustrates our design framework for the food justice dialog, where researchers, teachers, and computer scientists collaboratively engage in six phases: Rubric Development, Annotation, Model Building, Adaptive Guidance Design, Classroom Runs with Instructions, and Data Analysis & Revision. Each partner contributed varying levels of expertise across these phases (indicated by darker colors showing more intensive involvement), with researchers leading theoretical and analytical work, teachers providing pedagogical grounding and classroom implementation, and computer scientists driving technical development.

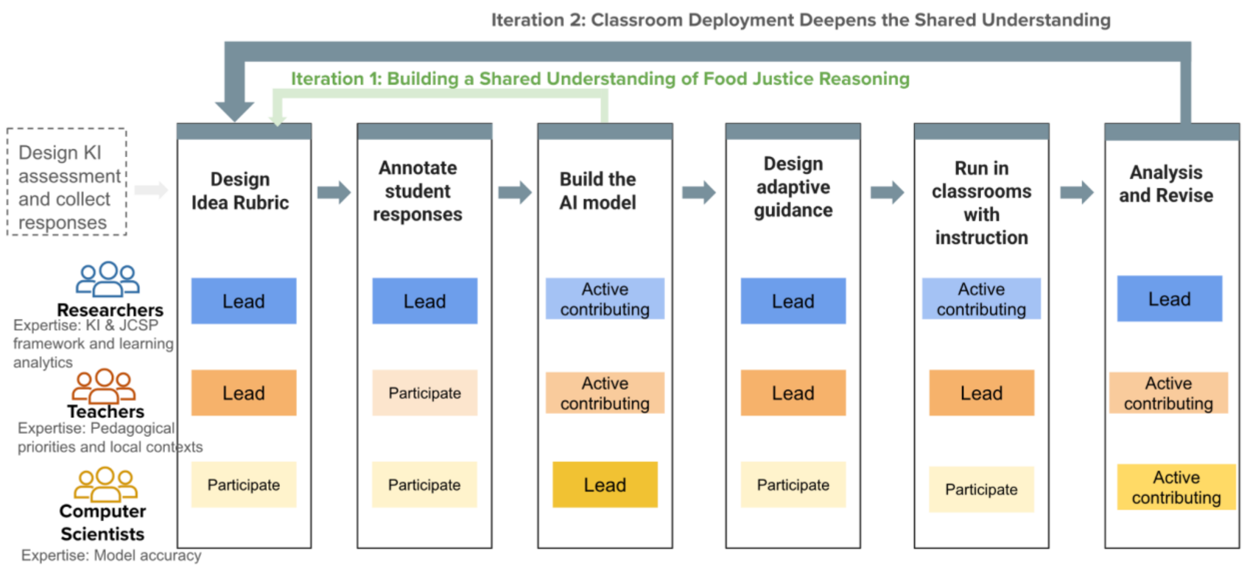

Figure 2. RPP Partner roles across design framework. Darker colors indicate more involvement.

Teachers (N = 8, 4 leading including Ben and Amy, 4 participating) provide essential grounding in local classrooms. During rubric development, leading teachers contributed in 1:1 meetings with the lead researcher (me) and full RPP meetings to identify pedagogically important ideas that students typically express and flag language patterns indicating emerging understanding. During adaptive guidance design, the leading teachers contributed to designing and revising guidance reflecting authentic teaching practices. During classroom runs, all 8 teachers taught the School Garden and Food Justice Unit with embedded AI dialogs. In the revision cycle, teachers attended RPP review meetings and workshops to identify which detected ideas would inform their instruction versus which created additional burden. The leading teachers also used the accuracy and frequency model output to adjust instruction. The teachers reviewed guidance and made suggestions after the rubric was finalized.

Computer scientists (N = 3) translate pedagogical goals into technical specifications. During Model Building phases (Iterations 1-3), computer scientists optimized architectures, implemented training procedures, and analyzed performance metrics. In full RPP meetings, they communicated feasibility constraints and possibilities, helping establish realistic expectations.

Educational researchers (N = 7) synthesize insights across partners using theoretical frameworks and learning analytics. The lead researcher (me) coordinated day-to-day partnership activities, conducting 1:1 meetings with teachers during Rubric Development and Adaptive Guidance Design, leading Annotation meetings every 2 weeks with one additional researcher over 3 months (Iterations 1 and 3), and managing data collection during Classroom Runs. The broader research team (N=5) contributed analytical expertise and review-level feedback through the full RPP review meetings across all phases. Researchers facilitated translation between pedagogical priorities and technical capabilities, ensuring that model development remained aligned with learning goals. For example, researchers leveraged the KI framework to guide keep/merge/drop decisions toward ideas that advance linking and systemic reasoning. They also applied the KI scoring rubric to evaluate learning gains from initial to revised responses to monitor breadth and balance across iterations.

4 Curriculum and technology contexts

4.1 Curriculum Context

The food justice dialog is embedded in the School Garden Web-based Inquiry Science Environment (WISE) unit, which integrates photosynthesis and ecosystems science with food justice investigation (Figure 3). The unit is localized around a school garden anchoring phenomenon and includes: (1) models and processes of photosynthesis, cellular respiration, and ecosystems aligned with NGSS MS-LS1 and MS-LS2; (2) a school garden modeling and design challenge applying plant photosynthesis processes to real constraints (drought, shade, nutrition); (3) a garden ecosystem connecting photosynthesis and cellular respiration to fresh food access; and (4) a food access investigation that connects the school garden to broader community contexts by examining structural inequities through exploration of local fresh food access patterns and historical redlining near three middle school campuses.

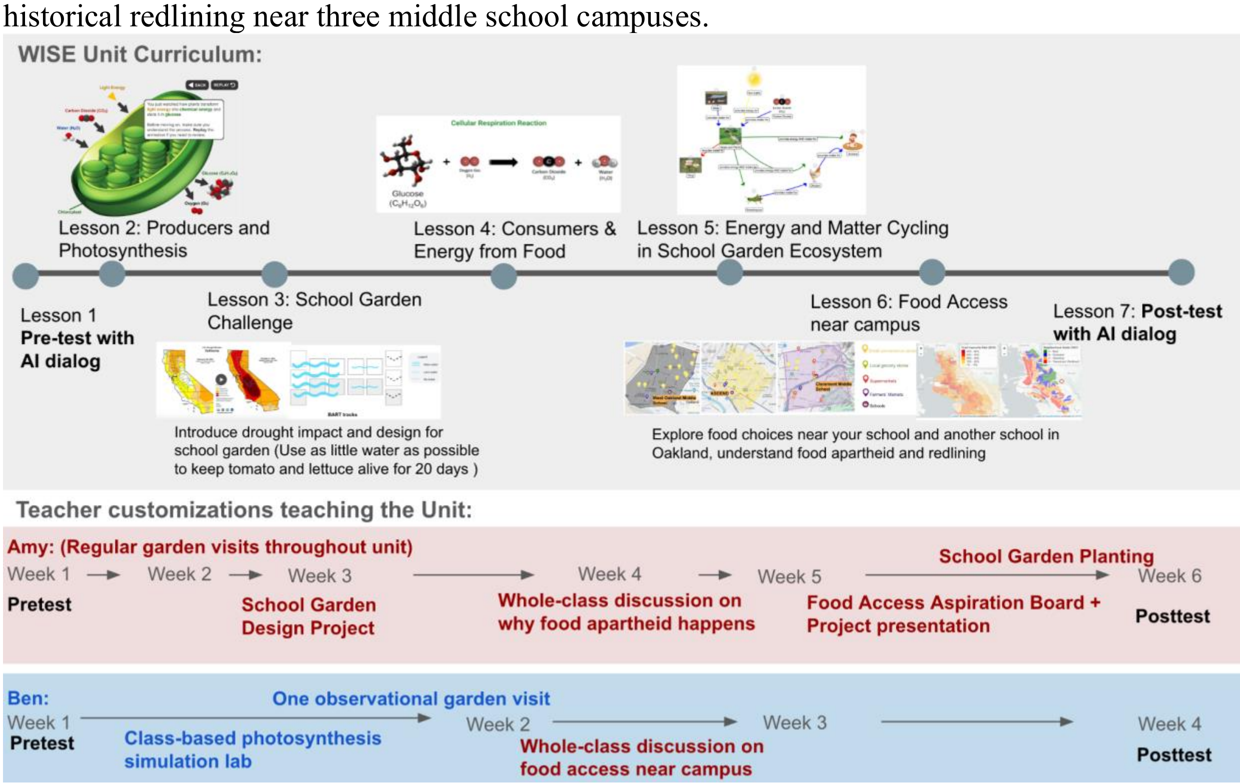

Figure 3. The School Garden and Food Justice unit plus teacher customizations

The WISE unit was designed by the RPP to enact the KI+JCSP design principles for food justice instruction as a shared instructional baseline for all participating classrooms. Making food justice accessible was enacted at two levels. At the community level, the school garden challenge in Lesson 3 asked students to connect photosynthesis and plant growth to students' own school community. The food access map activity (Lesson 6.4) asked students to explore the types and distribution of grocery stores near their school campuses using real neighborhood data. Prior to unit customization, I conducted pre-implementation meetings with teachers at each partner school to identify the grocery stores, markets, and food resources students recognized and used in their own neighborhoods, as well as nearby neighborhoods that could serve as contrast cases. These meetings informed school-specific food access maps that reflected each community's actual landscape, grounding food justice reasoning in what students could directly observe and recognize from their own lives rather than in generic neighborhood data. At the structural level, the redlining activity (Lesson 6.6) introduced historical redlining policy as a cause of current food access patterns, using side-by-side maps of food insecurity rates and 1937 neighborhood grades to make the connection between past policy and present inequity visible through data students could examine themselves. This made structural reasoning accessible through students' own neighborhood contexts rather than through abstract policy description. Making food justice thinking visible was enacted through explanation activities embedded throughout the unit, including the food justice AI dialog at pretest and posttest, which required students to articulate and link their ideas with evidence across two rounds of adaptive guidance. Helping students learn from others was enacted through peer discussion structures built into the activities, and promoting autonomy was enacted through the action planning activity (Lesson 6.7). This curriculum established what the dialog could build on across all classrooms; teacher customizations, examined in Phase 2, determined how deeply students engaged with the unit and which ideas integrated into their reasoning.

The food justice AI dialog (Figure 1) asks students to respond to the question: "Do you think all people have equal access to grow or buy fresh fruits and vegetables? Explain the reason why you think people do or don't have equal access." Students write an initial explanation, receive two rounds of adaptive guidance based on the ideas the NLP model detects in their response, then revise their explanation. The dialog appeared at pretest (before instruction) and posttest (after instruction).

4.2 NLP models

KI scoring rubric. Student learning was assessed using the five-level KI scoring rubric (Table 3), which measures integrated understanding by rewarding valid links among ideas rather than penalizing misconceptions. The rubric captures students' progression from off-topic responses (KI 1) and vague claims (KI 2) through single-factor explanations (KI 3) and two-factor links (KI 4) to multi-factor reasoning integrating socioeconomic, geographic, and systemic dimensions (KI 5).

Table 3
Rubric to the Food Justice KI assessment "Do you think all people have equal access to grow or buy fresh fruits and vegetables? Explain the reason why you think people do or don't have equal access."

KI Score	Criteria	Student Response Examples (detected ideas in the parentheses)
1	off-topic	"I'm not sure."

2	Vague or inaccurate ideas	"It depends on your race." (Vague race idea) "I think everyone can go to stores and choose what they like to eat." (Inaccurate claim)
3	Have one target idea	"Some People do not have the same access to fresh fruit or vegetables because some live far away from a nearby grocery store." (distance)
4	Link 2 target ideas	'Not all the people have the same access to fresh food because of how good the income is around the area so if you live in a low income area you would need to take the bus or walk several hours to the nearest supermarket." (SES + distance)
5	Link 3+ target ideas	"I think not everyone does because some of the community is in a food desert which mean that they don't have much access to fresh supermarket and there is only fast food because it is cheaper. And most times it is people of color or poor people living in those communities." (food desert + availability of supermarkets + fast food + food price + SES + race)

Idea detection model. The NLP idea detection model identifies food justice ideas in student responses, including vague and descriptive ideas (it depends on where you live), community-level evidence (distance to grocery stores, socioeconomic status, supermarket availability) and structural discrimination (redlining, food apartheid, race/ethnicity correlation). The idea rubric was developed and refined across two design iterations; the final rubric and model performance are reported in Section 4 (RQ1). For the focal classroom analyses (RQ2), both teacher's classrooms used the final idea detection model. Model performance was evaluated using micro-averaged F1 scores across idea categories.

To train the AI models, two researchers labeled student responses using INCEpTION (Klie et al., 2018), an open-source annotation platform, where they highlighted phrases representing specific ideas and assigned corresponding labels from the idea rubric. Two researchers conducted collaborative annotation in two rounds using 20% of responses to establish reliability and refine their shared understanding, achieving strong inter-rater agreement (Cohen's Kappa > 0.72) before independently annotating the remaining 80% of responses. This annotated dataset served as training data for the AI model.

Computer scientists from the RPP fine-tuned two transfer-learning models, one for KI scoring and one for idea detection, on DeBERTa-v3 (He et al., 2021), using a sequence-classification head for KI scoring and a token-labeling head for idea detection, following prior work on token classification in educational NLP (Riordan, Cahill, et al., 2020; Schulz et al., 2019). Hyperparameters including epochs, learning rate and schedule, warmup, and weight decay were selected via 10-fold cross-validation.

5 Methods

5.1 Participants

Student participants across both phases came from middle school science classrooms in the western United States. In Phase 1, two cohorts of students from five partner schools participated. The first cohort (N = 798) provided written responses to the KI assessment to build and train the initial NLP models. The second cohort (N = 869) interacted with the deployed dialog across benchmark assessments and classroom runs. Their dialogs were used to iterate the Food Justice dialog. Focal classroom analyses in Phase 2 included Ben's 6th-grade classrooms (N = 54) and Amy's 8th-grade classrooms (N = 53), both from the second cohort. These students represent a different cohort from the photosynthesis and ecosystems participants in Chapters 2 and 3, though they attended the same partner schools and shared similar demographic backgrounds.

5.2 Data Sources

For RQ1, data included student written responses from both cohorts (N = 798 and N = 869) collected across benchmark assessments and classroom runs, RPP meeting notes and workshop records documenting design decisions across both iterations, and model performance metrics (F1, precision, recall per idea) generated by computer scientists at each iteration. For RQ2, quantitative data included student dialog data from Ben's and Amy's classrooms at pretest and posttest, both using the final iterated dialog. Qualitative data included pre-implementation meeting transcripts (Ben: 1 meeting; Amy: 3 meetings), classroom observation fieldnotes collected by me across the majority of instructional sessions for both teachers, and post-implementation interviews conducted after unit completion.

5.3 Data Analysis

For Phase 1 (Dialog development), the model performance was analyzed using per-idea and overall micro-averaged F1 scores, precision, and recall across two iterations. I reported KI gains and the KI-idea heat map from the benchmarked assessments from the last iteration.

For Phase 2 (Teacher customizations), qualitative analysis of pre-implementation meeting transcripts, observation fieldnotes, and post-implementation interview data used open coding (Strauss & Corbin, 1998) to identify patterns in teachers' pedagogical goals and instructional moves. I organized these patterns using two complementary analytical lenses. First, JCSP approaches (Morales-Doyle, 2018) characterized each teacher's overall pedagogical orientation toward food justice instruction: whether they framed food justice as injected relevance, everyday relevance, or a problem-posing justice-centered approach. Second, the four KI+JCSP design principles for the food justice instruction (Table 1 in Section 2) characterized teachers' specific instructional moves and customizations, capturing how each teacher enacted making SJSI accessible, making SJSI thinking visible, helping students learn from others, and promoting autonomy within their overall pedagogical orientation.

Quantitative analysis of student dialog data used cumulative link mixed models (CLMMs) to examine KI score changes, appropriate for ordinal outcome data, with fixed effects for dialog stage (initial vs. revised explanation), timepoint (pretest vs. posttest), and their interaction, and random intercepts for student. I report raw mean KI scores in tables for interpretability; CLMM estimates reflect statistical inference controlling for student-level random effects and will differ from raw means accordingly. Wilcoxon signed-rank tests examined whether the dialog helped students add new ideas beyond their initial responses. McNemar's tests with Benjamini-Hochberg correction examined which specific ideas students incorporated during dialog interaction.

6 Results

6.1 RQ1: Iterative Design of a Food Justice Dialog

Before the RPP could build a model to detect food justice ideas, the partnership had to agree on what those ideas were, whose language counted, and which ideas to scaffold in the school garden and food access reasoning context. That shared understanding was the product of Phase 1. Two iterations of design, each representing a complete design-implementation-revision cycle, produced a 22-idea rubric grounded in authentic student language and an NLP model achieving an overall micro-averaged F1 of 0.73 across 19 detectable ideas. This section reports how those two iterations unfolded and what the partnership learned from each.

6.1.1 Iteration 1: Building a Shared Understanding of Food Justice Reasoning

Two researchers developed the initial idea rubric through four rounds of open coding on 200 randomly selected student responses from an initial pilot from two partner schools. This inductive process identified recurring themes organized into five broad categories: economic factors, geographic factors, transportation, systemic inequities, and individual choice. Iterative refinement across four rounds reached satisfactory inter-rater reliability (Cohen's Kappa > 0.80), resulting in an initial 20-idea rubric. The full RPP team (4 teachers including Amy and Ben, 7 researchers, and 3 computer scientists) reviewed and agreed on the rubric and idea definitions.

Using this rubric, the two researchers annotated 798 student responses collected across benchmark assessments and classroom runs in partner schools. Computer scientists trained an initial NLP model on this annotated dataset, achieving an overall micro-averaged F1 score of 0.71 with 11 reliably detectable ideas. The model performed well on high-frequency ideas: SES (F1 = 0.81) and "not all equal access" (F1 = 0.82). However, several pedagogically important ideas showed near-zero detection accuracy due to low frequency and linguistic variability, including "entrepreneurs build fast food in food deserts" (F1 = 0.00), "lack of transportation" (F1 = 0.00), and "systemic discrimination policy" (F1 = 0.00).

The full RPP team then reviewed per-idea F1 scores, precision, recall, and label frequency, making collaborative keep/merge/drop decisions grounded in two criteria: model performance and instructional value. This dual-criteria approach surfaced productive tension between partners. Computer scientists advocated for removing ideas the model could not reliably detect; teachers advocated for retaining ideas that mattered for classroom instruction, reasoning that what students could not yet express clearly was precisely what instruction should support.

Three types of decisions resulted from this negotiation. Teachers suggested removing very infrequent ideas with little instructional value, such as "fresh food is individual choice." Teachers and researchers merged conceptually redundant categories, such as folding "availability of nearby convenience stores" into "availability of fast food places." Most importantly, teachers and computer scientists agreed to retain low-frequency but pedagogically central ideas when precision was sufficient, notably "social transformation" and "fast food is cheaper than fresh produce."

The retention of "social transformation" despite low model accuracy illustrates what the RPP was really building in RQ1. In a review meeting, Amy argued that this idea indexes students' reasoning as transformative intellectuals (Morales-Doyle, 2017), a core JCSP goal, even when it appears rarely and in linguistically varied forms. Dropping it would mean the dialog could never recognize or respond to an important justice-centered reasoning students produced. Keeping it meant committing to teach toward it even when the model could not fully see it. This decision embedded a shared value into the rubric: food justice reasoning matters most at its structural and transformative edges, and the partnership would not let technical limitations define the ceiling of what students were expected to think.

After consolidating the rubric to 15 ideas, researchers and teachers designed two rounds of adaptive guidance for each idea, drawing on teachers' in-class responses to student reasoning. Computer scientists retrained the model on the same 798 annotated responses using the revised rubric. The final Iteration 1 model achieved an overall F1 of 0.74 with 15 detectable ideas. Core categories remained stable: SES (F1 = 0.81), "not all equal" (F1 = 0.81), "far away from grocery" (F1 = 0.77). Pedagogically important ideas improved: photosynthesis-related ideas (F1: 0.74 to 0.77), food desert (F1: 0.61 to 0.63), systemic discrimination policy (F1: 0.00 to 0.43). Following model refinement, teachers and researchers also revised the WISE unit to add evidence-based reasoning contexts: a food access map showing types and distribution of grocery stores near partner schools, and background on redlining with an interactive redlining map. The Iteration 1 model was then deployed in classrooms.

6.1.2 Iteration 2: Classroom Deployment Deepens the Shared Understanding

A new cohort of 869 students from the same partner schools used the Iteration 1 dialog across benchmark assessments without instruction and classroom runs. These same 869 dialogs served dual purposes: they provided evidence of the dialog's effectiveness and became the training data for the Iteration 2 model. This dual role reflects a core feature of iterative RPP design: authentic classroom data simultaneously validates the current model and informs the next one.

Classroom deployment revealed reasoning patterns that the 798-response training data had not captured. Students' multi-turn dialog responses showed a progression the model had not been designed to recognize: beginning with vague claims such as "not everyone has the same access," then specifying scientific phenomenon (e.g., drought affecting plant photosynthesis), human impacts, and observations about nearby grocery store types and numbers after adaptive guidance. This progression was only visible in multi-turn sequences, not in isolated responses. Analyzing how students' reasoning developed across dialog turns, rather than treating each response as a discrete data point, was essential for capturing how the dialog actually worked in classrooms.

Three researchers (me and two pre-service teacher researchers) first annotated 200 dialogs out of all 869 dialogs using the Iteration 1 rubric, then identified ideas that needed revision based on patterns observed in authentic classroom reasoning. New distinctions emerged from students' own language that the initial rubric had not anticipated. The differences between “systemic discrimination,” “individual discrimination,” “vague discrimination,” and “observed race/ethnicity correlation as evidence” reflected meaningfully distinct reasoning moves that the 15-idea rubric had collapsed into a single category. Recognizing these distinctions required the researchers to return to the RPP with new questions: were these the same idea or different ones? Did teachers see these distinctions in their classrooms? Did they matter instructionally? These conversations deepened the partnership's shared understanding of food justice reasoning, expanding the rubric from 15 to 22 ideas (Table 4) grounded in the language students actually used. After reaching inter-rater reliability with the new rubric, three researchers then annotated the rest of the 869 dialogs over three months.

Computer scientists retrained the model on the 869 annotated classroom dialogs. The Iteration 2 model achieved an overall F1 of 0.73 across 19 detectable ideas (Table 4, detailed idea description in Appendix Table 5). Performance strengthened on core learning targets: SES (F1 = 0.83), distance barriers (F1 = 0.80). Idea 8, factors affecting plant photosynthesis and growth, remained stable and reliably detectable (F1 = 0.81), which confirmed that the unit's core science-to-food-justice integration was working as designed: students were drawing on plant growth conditions like drought and shade tolerance to explain why some communities can grow or access fresh food and others cannot. "Availability of gardens" improved substantially (F1: 0.75) and systemic discrimination emerged as a reliably detectable category (F1: 0.43 to 0.72).

Table 4. 22-idea rubric and per-idea accuracy after Iteration 2

Original rubric at Iteration 1	F1 score	Revised rubric at Iteration 2	F1 score
Community-based evidence collected from observations, data, and instruction			

Idea 4: Availability of fast food	0.70	Idea 4: Availability of fast food	0.71
Idea 5: availability of grocery stores/supermarket/farmers market	0.62	Idea 5: availability of fresh produce (e.g., grocery stores/supermarkets)	0.61
-	-	Idea 7: availability of convenience stores	0.12
-	-	Idea 6: availability of community gardens	0.75
Idea 9: Far away from grocery/far drive	0.77	Idea 9: Far away from grocery/far drive	0.80
Idea 10: Lack of transportation	0.00	Idea 10: Lack of transportation	0.63
Idea 13: SES	0.81	Idea 13: SES	0.83
Idea 15: Fresh food is expensive	0.63	Idea 15: Fresh food is expensive	0.67
Idea 16: Fast food is cheap	0.28	Idea 16: Fast food is cheap	0.36
Science mechanism			
Idea 8: Factors affecting photosynthesis of growing plants	0.77	Idea 8: Factors affecting photosynthesis of growing plants	0.81
Misapplied science reasoning			
-	-	Idea 22: Pre-existing health conditions or diet impact on health	0.62
-	-	Idea 23: nutrition value	0.00
-	-	Idea 24: resources to grow plants (e.g., seeds, tools, knowledge)	0.40
Systemic Injustice			
Idea 14: Systemic Discrimination	0.43	Idea 14: Systemic Discrimination policies (redlining)	0.72
Idea 18: Race/ethnicity correlation	0.46	Idea 18: Race/ethnicity correlation	0.58
-	-	Idea 26: Culturally relevant food	0.17
Idea 17: Social transformation	0.17	Idea 17: Social transformation	0.25
Idea 11: Food desert	0.63	Idea 11: Food apartheid	0.69
-	-	Idea 12: Companies/Government put fast food in food desert	0.70
Vague/Descriptive/Non-target			
Idea 1: All equal access	0.73	Idea 1: All equal access	0.72

Idea 1b: Not all equal	0.81	Idea 1b: Not all equal	0.82
Idea 3: Vague location or country/city	0.62	Idea 3: Vague location or country/city	0.64
-	-	Idea 25: Vague discrimination (e.g., it's all about race)	0.00

Table 4 presents the final Iteration 2 rubric organized into four categories that reflect the integrated KI and JCSP framework underlying the dialog's design. Target social justice ideas are divided into two levels. Community-based evidence ideas capture what students can observe, discover, or learn from neighborhood data and classroom instruction. This tier includes the availability of grocery stores, supermarkets, and community gardens in their neighborhoods (Ideas 4, 5, 6); geographic barriers such as distance to grocery stores and lack of transportation (Ideas 9, 10); and economic factors including socioeconomic status and food prices (Ideas 13, 15, 16). Two ideas in this tier deserve specific definition because they are conceptually related but capture distinct aspects of food access. Distance to grocery stores (Idea 9) refers to the geographic barrier students face in reaching fresh food, such as how far they must walk, take a bus, or be driven to find a store that sells fresh produce. Supermarket availability (Idea 7) refers to the number and distribution of stores selling fresh produce in a given neighborhood, which may include large supermarkets or smaller local and community-owned grocery stores. A neighborhood can have convenience stores nearby but still have few that carry fresh produce, or have stores that are culturally relevant but not counted in standard supermarket data. Socioeconomic status (SES, Idea 13) is classified as community-based evidence rather than structural because students can directly observe economic disparity in their own neighborhoods and communities. They see that some families have more resources than others without necessarily reasoning about the structural and historical policies that produced those inequities. When students name SES as a factor in food access, they are drawing on observable community conditions; when they name redlining or food apartheid, they are reasoning about the structural causes of those conditions. Community-based evidence ideas are accessible through dialog scaffolding of students' own observations and the WISE unit's food access map activity, and the model detects most of them reliably. Structural injustice ideas capture systemic reasoning about the historical and current policies and practices: food apartheid (Idea 11), corporate and government decisions that concentrate fast food in underserved neighborhoods (Idea 12), systemic discrimination policies such as redlining (Idea 14), race and ethnicity injustice (Idea 18), and social transformation reasoning in which students position themselves as agents of change (Idea 17). These ideas are central to JCSP goals and reflect the most sophisticated food justice reasoning students can produce, yet they are also the hardest for the model to detect reliably.

Some ideas remained difficult to detect reliably: social transformation (F1 = 0.25) and observed race/ethnicity correlation (F1 = 0.58). These persistent limitations reveal the nature of justice-centered reasoning: students' expressions of structural critique are linguistically diverse, contextually embedded, and emotionally situated in ways that resist standardized detection. The RPP's decision in Iteration 1 to retain these ideas despite low accuracy proved consequential here. Because the rubric still included them, the partnership could see clearly where the model's ceiling was and what that meant for instruction. Shared understanding of the model's limits became as important as shared understanding of what the model could detect: it identified where teacher presence, relationship, and judgment would need to do what the dialog could not.

The heatmap analysis of students' revised responses from Iteration 2 (N = 869, Figure 4) validated the rubric design by showing that ideas distributed meaningfully across KI score levels. At KI 2, broad claims dominated: "all people have equal access" and "not everyone has equal access" appeared frequently, with few references to mechanisms. By KI 3, attention shifted toward structural explanations, with "not equal access" and "SES" most frequent and notable growth in "location" and "supermarkets," indicating emerging attention to neighborhood-level evidence. At KI 4, references to local food stores became more salient, with "fast food" and "supermarkets" increasing alongside continued mention of "not equal access" and "SES." At KI 5, students pulled evidence from multiple factors while bringing a systemic lens, with higher shares for "race/ethnicity," "food desert," and price-related ideas such as "fresh food is expensive" and "fast food is cheaper." This progression from broad claims toward multi-factor reasoning that integrates socioeconomic, spatial, and systemic dimensions of food access confirms that higher KI scores correspond to more sophisticated reasoning that the model can identify, validating both the rubric design and the detection model.

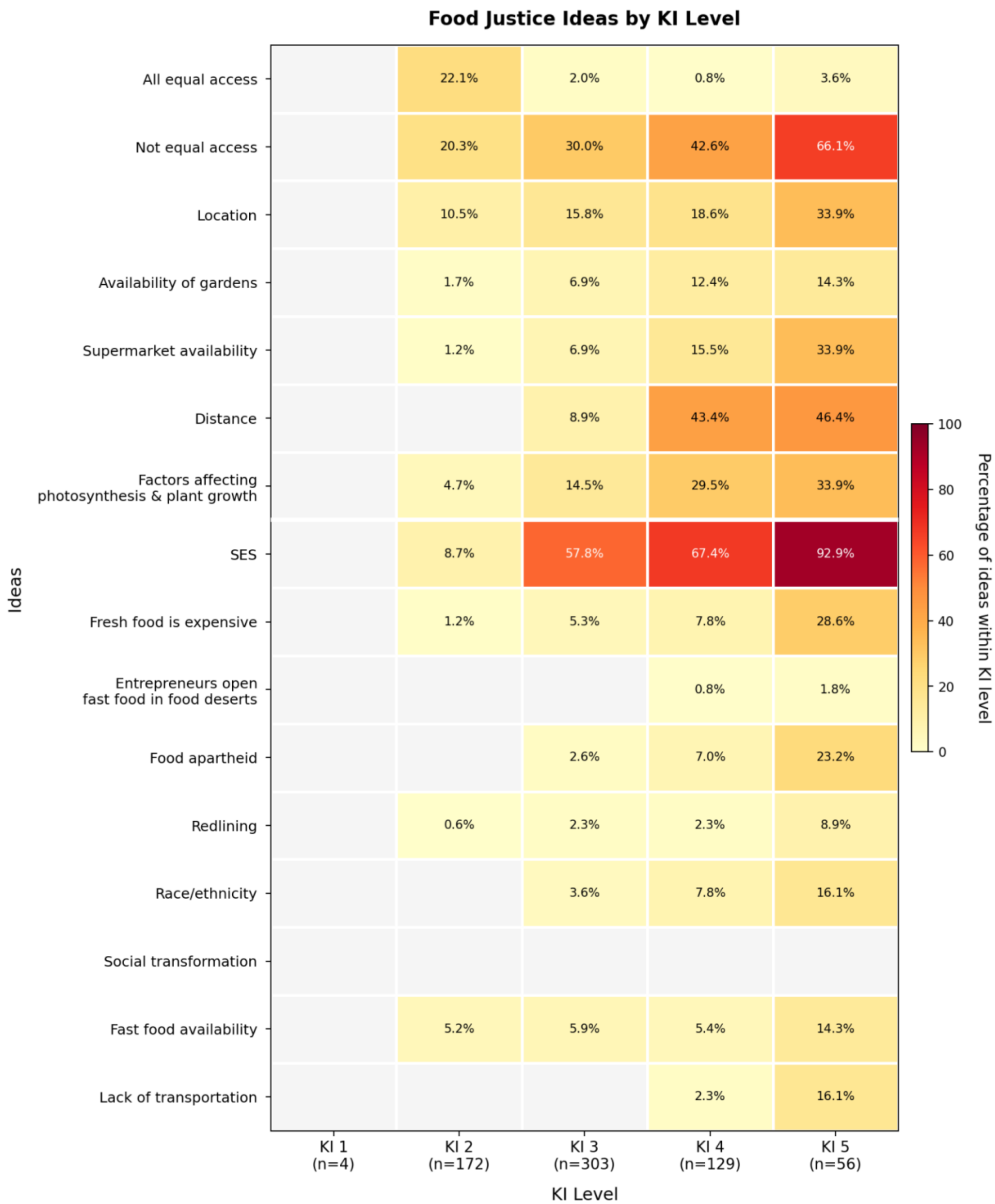

Figure 4. Idea heatmap across KI levels

6.1.3 Summary

Two iterations of RPP design produced more than a functional NLP model. The RPP built a shared understanding of what food justice reasoning looks like in middle schoolers' words, encoded in a 22-idea rubric that captures reasoning from community-level observation to structural reasoning to transformative action. Three insights emerged from this process. First, teachers' pedagogical priorities evolved when confronted with authentic student data: distinctions that seemed peripheral in benchmark responses became central after classroom deployment revealed how students actually reasoned across multi-turn dialog sequences. Second, the persistent tension between model performance and instructional value, particularly for race, discrimination, and social transformation ideas, produced shared understanding not only of what the model could do but of where teacher instruction would need to carry the work. Third, the dual role of classroom data in simultaneously validating the current model and training the next one demonstrates that iterative design requires authentic classroom implementation, not just benchmark data, to capture the full range of student reasoning about social justice science issues.

This shared understanding traveled into classrooms with Ben and Amy, but each teacher enacted it through a different pedagogical orientation. Prior to examining their implementations, it is worth noting that the dialog itself produced meaningful KI gains even in the absence of instruction: among students completing the Iteration 2 dialog as a benchmark assessment without embedded classroom instruction ($N = 559$), revised explanations were scored significantly higher than initial explanations ($V = 1178.0$, $p = .005$), with a mean gain of 0.71 KI levels. This establishes a baseline: the dialog works on its own. The question Study 2 takes up is what becomes possible when teachers bring their own knowledge, relationships, and pedagogical orientations to the food justice dialog and instruction.

6.2 RQ2: Teacher customizations of food justice instruction with AI dialogs

Study 1 established the dialog's technical reliability and the shared rubric that encodes what food justice reasoning looks like in middle schoolers' words. Study 2 examines what two teachers with different pedagogical orientations toward food justice accomplished when they brought that dialog into their classrooms.

6.2.1 Making Food Justice Accessible and Proximate: Dialog and Instruction in Ben's Classroom

Ben's Pedagogical Orientation

Ben approached food justice instruction by making SJSI accessible and personally proximate for his 6th graders, which is the first of the KI+JCSP design principles (Table 1). Ben's instruction centered on bringing food justice within reach of students who were encountering formal science instruction for the first time, grounding scientific concepts in what was immediately observable and personally meaningful in their communities.

Ben was explicit and theoretically grounded about this choice. During a pre-implementation meeting, he articulated his approach by drawing directly on Morales-Doyle's (2018) JCSP framework and approaches to make science accessible, which he had studied in his master's program: "We read some of Morales-Doyle's work and there's two ways of doing social justice teaching. You have the project-based where students are solving real community-based problems. The other model is the injected relevance model where you connect something that is hyper local to students and involves social justice to sort of grab their interest. And so I think that is how I am going to teach food justice here." This quote is notable not only for what it describes but for what it demonstrates: Ben was not defaulting to a shallow engagement approach. He was making a principled pedagogical decision, aware of the full range of JCSP approaches, and choosing the one he judged most appropriate for his students' developmental stage and context.

This judgment was grounded in his knowledge of his students. Ben's 6th graders were receiving their first formal science instruction, and many carried responsibilities at home that shaped what food justice meant to them personally: "A lot of my students are their parents' point of contact for the English-speaking world. Giving them those resources about food banks, they're like, 'Hey, if my family needs food, I know where I can go to get that from my school or my community.'" Ben saw empowerment at the proximal, accessible, and immediate scale as the appropriate goal for students at this age and in this context: "I think at their age, designing an action project to change food access is tricky. Because they're like 11, 12 years old. They know this is bad but they don't have a concrete plan. So I think they see ways to help their own community through the school garden where we can give our friends and family fruits and vegetables. We can help people that are immediately near to us."

This orientation shaped which KI+JCSP principles Ben enacted most fully. His customizations centered on making SJSI accessible through personal proximity and helping students learn from others through guided discovery, bringing food justice reasoning within reach of students who had not yet developed structural vocabulary to analyze it at a systemic level.

Ben's Instructional Activities

Ben implemented the WISE unit over approximately 12 class periods across four weeks, conducting one pre-implementation meeting with the researcher to review the unit and dialog guidance and adding one teacher-designed activity beyond the WISE content. His instructional moves consistently enacted two KI+JCSP principles: making SJSI accessible and making thinking visible.

Making SJSI Accessible and Thinking Visible. Ben's instruction consistently translated potentially abstract food justice content into what was personally observable and proximate for his students, while simultaneously making student thinking visible through teacher-mediated whole-class synthesis. The food access map activity in Lesson 6.4 (Figure 5) illustrates both principles working together. Before students explored the maps, Ben asked "What does food insecurity mean?" and wrote students' responses on the board: "less grocery stores," "no food," "cannot secure next meal," "only eat junk food." This activated prior knowledge and made existing student thinking visible before new content was introduced. As students explored the maps, Ben asked "What's a convenience store?" when students initially interpreted the abundance of yellow dots near their school as a sign of good food access. Students began discussing how convenience stores sell packaged food rather than fresh produce, arriving at the insight themselves that quantity does not equal quality. After exploration, Ben synthesized the pattern students had collectively noticed: their school's neighborhood had many convenience stores but few supermarkets, while other neighborhoods had more access to fresh produce through community-owned grocery stores. This teacher-mediated synthesis grounded students' food justice reasoning in community-based evidence ideas about store availability and supermarket access (Ideas 4, 5, 7), making neighborhood-level patterns visible to the whole class before students moved to individual WISE responses.

Figure 5. Food access map activity on WISE unit Lesson 6.4

Ben applied the same accessibility approach during the redlining activity in Lesson 6.6 (Figure 6), where students examined side-by-side maps of food insecurity rates and 1937 neighborhood grades near their school campuses. He began by asking "What does red mean?" When students responded "hazardous," he clarified: "Well it actually means hazardous for mortgage support. But many people just assume it is hazardous and not good to live in." He then connected historical policy to students' lived experience through questions: "Do people move far away from their families? How might historical policy still impact us today? If your parents don't have much money, would you also not be able to move to nicer neighborhoods?" These questions helped students reason about structural inequity from a starting point they already understood personally, making the redlining map's hazardous neighborhood markings meaningful as evidence rather than as distant history. Ben acknowledged the structural pattern students noticed: "As [student K] mentioned, we noticed that there are some racism going on here when they marked those red places as hazardous." This acknowledgment reflected Ben's approach throughout the unit: he made structural ideas accessible by connecting them to what students could already observe, using teacher-mediated synthesis to name patterns rather than structured peer dialogue to generate critique. This kept the structural content within reach of students who were encountering these ideas for the first time.

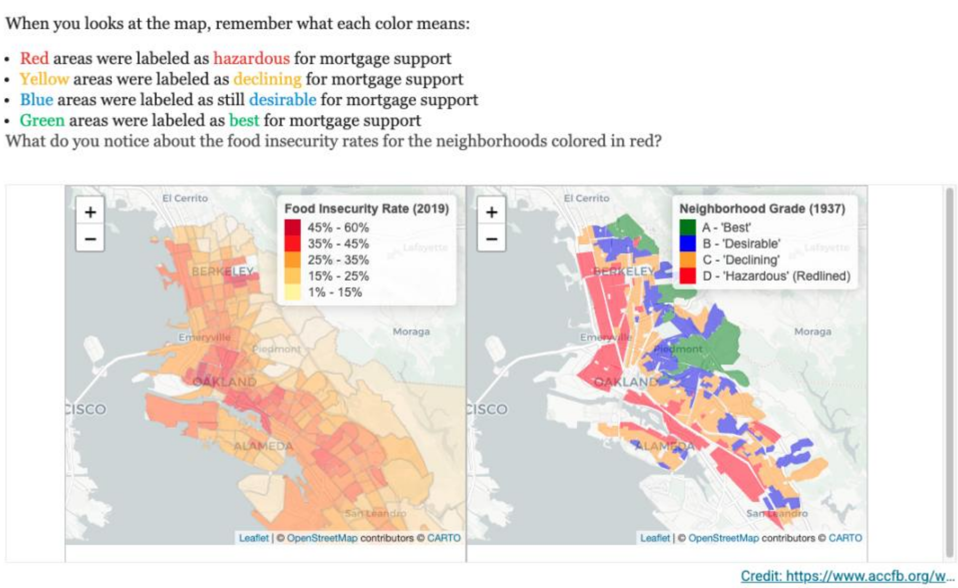

Figure 6. The food insecurity and redlining maps activity in Lesson 6.6

Making SJSI Thinking Visible. Ben's teacher-designed activity beyond the WISE unit enacted the making thinking visible principle at the action planning stage. During Lesson 6.7 (Develop Your Action Plan), he brought the brainstorming activity off the screen and onto the board: "Each table needs to come up with one idea and I'm gonna call on one of the students. I write all of their ideas up in the box." Students contributed ideas through Ben, who recorded and synthesized them collectively, making the full range of student thinking visible before students completed individual WISE responses. Students identified the school garden as a resource for growing fresh produce and noted the availability of food pantries and food banks near campus, ideas that aligned with Ben's empowerment framing.

A spontaneous KI-discovering moment near the end of the unit illustrated the cumulative effect of Ben's accessible, visible framing. After Ben mentioned the community food resources page in WISE, students began independently clicking links to food banks and taking notes, driven by personal relevance rather than assignment requirements. Ben reflected: "Giving them those resources, they're like, 'Hey, if my family needs food, I know where I can go.' Which is not something 12, 13-year-olds should be figuring out, but some of them started clicking on the links." This moment was not planned, but it was made possible by the proximal, personally relevant framing Ben had sustained throughout the unit.

Ben's instructional focus remained proximal and personal throughout: "getting them thinking more about eating fresh fruits and vegetables and that they have the ability to get them at their school, I hope they felt more empowered in that direction." He walked students through the WISE content on redlining and food deserts by connecting historical policy to what students could personally observe and relate to, using teacher synthesis to make structural patterns visible rather than peer dialogue to interrogate them. This was a principled choice grounded in his knowledge of his students' developmental stage and prior experience with science instruction. It determined which ideas were available in students' reasoning when they engaged with the dialog, and what the dialog could help them articulate and integrate.

Ben's Student Learning Outcomes

The pattern of ideas students incorporated through dialog reflects the ideas Ben's instruction made accessible. Students showed consistent and significant KI gains through dialog interaction at both pretest and posttest, and ideas that were not specifically scaffolded through explicit instructional activities were not significantly incorporated.

KI Gains. A CLMM revealed a significant main effect of dialog across both timepoints ($\beta = 2.48$, $SE = 0.47$, $z = 5.31$, $p < .001$), indicating that revised explanations were scored substantially higher than initial explanations. Dialog gains were consistent: mean initial KI 2.20 (SD = 0.84) to mean revised KI 3.02 (SD = 0.89) at pretest (+0.78); mean initial KI 2.48 (SD = 0.81) to mean revised KI 3.22 (SD = 0.90) at posttest (+0.76). The marginal main effect of instruction ($\beta = 0.79$, $p = .062$) suggests modest growth in students' unprompted knowledge across the unit, though this did not reach conventional significance. The absence of a significant Dialog × Timepoint interaction indicates the dialog was equally effective before and after instruction. Table 5 presents raw mean KI scores across timepoints.

Table 5. *KI Scores Across Timepoints in Ben's Classrooms (N = 54)*

Timepoint	Initial KI M (SD)	Revised KI M (SD)	KI Gain
Pretest	2.20 (0.84)	3.02 (0.89)	+0.78

Posttest	2.48 (0.81)	3.22 (0.90)	+0.76

Figure 7 presents the estimated marginal means from the CLMM. The estimated marginal means from the CLMM differ from raw means in Table 3 due to adjustment for student-level random effects; both are reported for transparency. In Ben's classrooms (left panel), the dialog produced consistent gains at both timepoints: estimated KI scores increased from 2.22 to 3.02 at pretest (+0.80) and from 2.45 to 3.17 at posttest (+0.72). The nearly parallel pattern between pretest and posttest bars reflects the non-significant Dialog × Timepoint interaction. The modest increase in initial KI scores from pretest (2.22) to posttest (2.45) is consistent with the marginal instruction effect, suggesting some growth in students' unprompted knowledge across the unit.

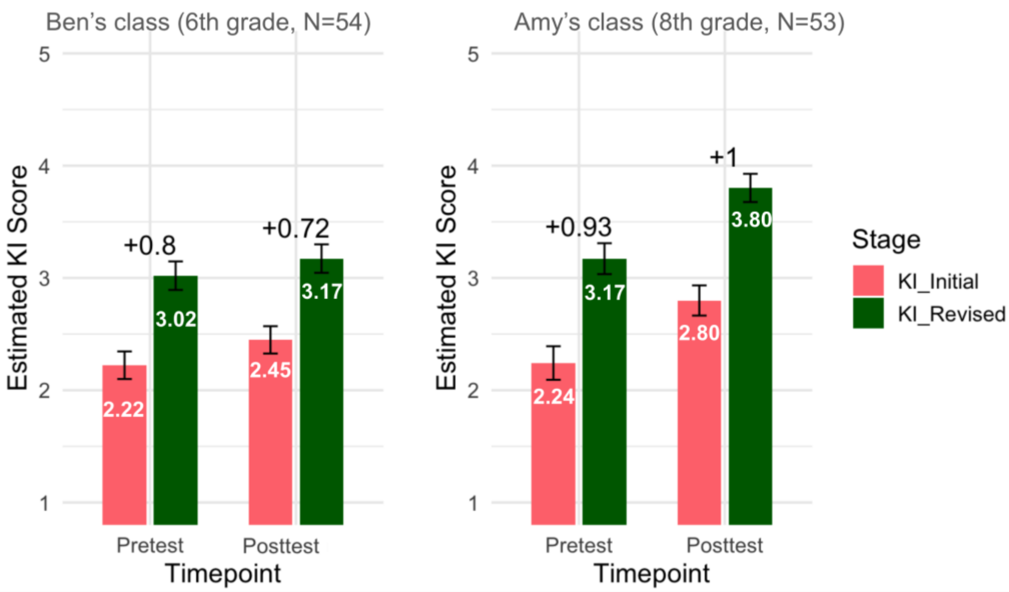

Figure 7. Estimated Marginal Means of KI Scores by Dialog Stage and Timepoint

Ideas Added During Dialog. Wilcoxon signed-rank tests confirmed the dialog's scaffolding function at both timepoints. At pretest ($N = 49$), students added an average of 1.00 new idea ($SD = 0.96$; $V = 528$, $p < .001$), including a mean of 0.65 target ideas ($SD = 0.75$; $V = 300$, $p < .001$). At posttest ($N = 54$), students added an average of 0.98 new ideas ($SD = 1.00$; $V = 528$, $p < .001$), including 0.59 target ideas ($SD = 0.69$; $V = 351$, $p < .001$). No significant difference in ideas added across timepoints (Total: $W = 1345$, $p = .881$; Normative: $W = 1366$, $p = .756$) indicates the dialog's scaffolding function remained stable throughout instruction, consistent with the non-significant instruction effect.

Specific Ideas Incorporated. McNemar's tests with Benjamini-Hochberg correction identified which specific ideas students incorporated through dialog interaction (Table 6). At pretest, before any instruction, students most frequently added ideas about unequal access as a claim (+14.3%, $p = .027$), distance to grocery stores (+20.4%, $p = .008$), and socioeconomic status (+28.6%, $p < .001$). These are community-based evidence ideas that students could draw on from everyday observation before instruction: they already knew from their own lives that some neighborhoods are farther from fresh food stores and that some families have fewer resources than others. The dialog elicited that existing community knowledge at pretest.

Table 6. Key ideas Added During Dialog (Initial to Revised Responses) in Ben's classrooms

Ideas	Time	N	**Initial %**	**Revised %**	**Change**	χ^2	**p (uncorr.)**	**p (BH)**
1b: Claim - Not Equal Access	Pretest	49	10.20%	24.50%	14.30%	5.44	0.0196	0.02744
	Posttest	54	13%	38.90%	25.90%	12.25	0.0005	0.00175
8: Factors affecting Photosynthesis	Pretest	49	0.0%	4.9%	4.9%	0.50	0.4795	0.4795
	Posttest	54	4.5%	9.1%	4.6%	0.50	0.4795	0.6393
9: Distance to Grocery Stores	Pretest	49	6.10%	26.50%	20.40%	8.33	0.0039	0.00805
	Posttest	54	18.50%	42.60%	24.10%	8.05	0.0046	0.00805
13: Socioeconomic Status	Pretest	49	22.40%	51%	28.60%	14	0.0002	0.0014
	Posttest	54	22.20%	29.60%	7.40%	1.33	0.2482	0.2482
11: Food apartheid	Pretest	49						
	Posttest	54	0	5%	5%	0.78	0.439	0.4392
7: Supermarket Availability	Pretest							
	Posttest	54	3.70%	15.10%	11.40%	4.02	0.0473	0.083517
14: Redlining Policy	Pretest							
	Posttest	54	0	7.2%	7.20%	1.09	0.336	0.420

At posttest, the pattern shifted in ways that reflect Ben's instruction. Students continued to incorporate distance to grocery stores (+24.1%, $p = .008$) and unequal access (+25.9%, $p < .001$), consistent with the accessibility framing Ben sustained throughout the unit. Supermarket availability also reached significance at posttest (+11.4%, $p = .047$), reflecting the specific instructional work of Ben's food access map discussion, in which he guided students to distinguish the number and type of fresh food stores across neighborhoods and synthesized those observations for the whole class. This is community-based evidence about how many stores selling fresh produce exist in a given neighborhood, and Ben's whole-class synthesis made it observable and nameable for students in a way that carried into their dialog responses. Socioeconomic status, by contrast, was no longer significant at posttest (+7.4%, $p = .248$): students entered the posttest dialog already mentioning SES in their initial responses at roughly the same rate as pretest, suggesting it had become part of their prior knowledge rather than an idea the dialog needed to elicit.

Structural ideas like food apartheid and redlining policy did not increase significantly from pretest to posttest. This pattern reflects the instructional emphasis of Ben's accessibility-focused approach. The dialog reliably helped students elaborate community-based evidence ideas that instruction had made personally observable. Ideas that require explicit vocabulary work to develop, such as redlining and food apartheid, were present in the WISE unit but were not the focus of Ben's instructional customizations, and did not significantly improve.

Idea 8, factors affecting plant photosynthesis and growth, showed modest but non-significant increases at both timepoints (pretest: +4.9%, $p = .480$; posttest: +4.6%, $p = .480$). Ben's fresh food map observations and school garden activities introduced the connection between plant growth and food access, and the dialog elicited this reasoning at a consistent level across both timepoints. The absence of significance reflects that plant-growth-to-food-access reasoning, while present in students' revised responses, was not a primary focus of Ben's accessibility instruction. Students drew on it modestly when the dialog prompted them, using science mechanisms to explain why some communities can grow or access fresh food, but did not carry it forward as a central evidence. This pattern is consistent with Ben's instructional focus on what was immediately observable: the gap between convenience stores and supermarkets, the distance to fresh food, the availability of community resources.

Student J's dialog interactions (Figure 8) illustrate how the dialog helped students add and integrate community-based evidence ideas that instruction had made available. At pretest, the dialog prompted J to go beyond an initial response about naming the factors like cost and seed availability, eliciting reasoning about distance and the absence of nearby stores. J's revision added these ideas but stayed at the level of personal observation. At posttest, the dialog's guidance produced richer responses because Ben's food access map discussion had stocked J's knowledge with new observational evidence: J's initial response already referenced chart and map data (in Lesson 6.4) about the distribution of food stores across neighborhoods before the dialog began. As the dialog asked J to explain why supermarket chains open fewer stores in some neighborhoods, J drew on the redlining map activity (Lesson 6.6) from the WISE unit, recalling that some areas were marked as hazardous, and connecting those markings to where grocery stores choose to locate. J did not use the word redlining, but the visual evidence from the map was present in J's reasoning. The dialog gave J a structured opportunity to articulate and integrate what the food access and redlining map activities had made available, even without explicit vocabulary for the structural idea underlying it.

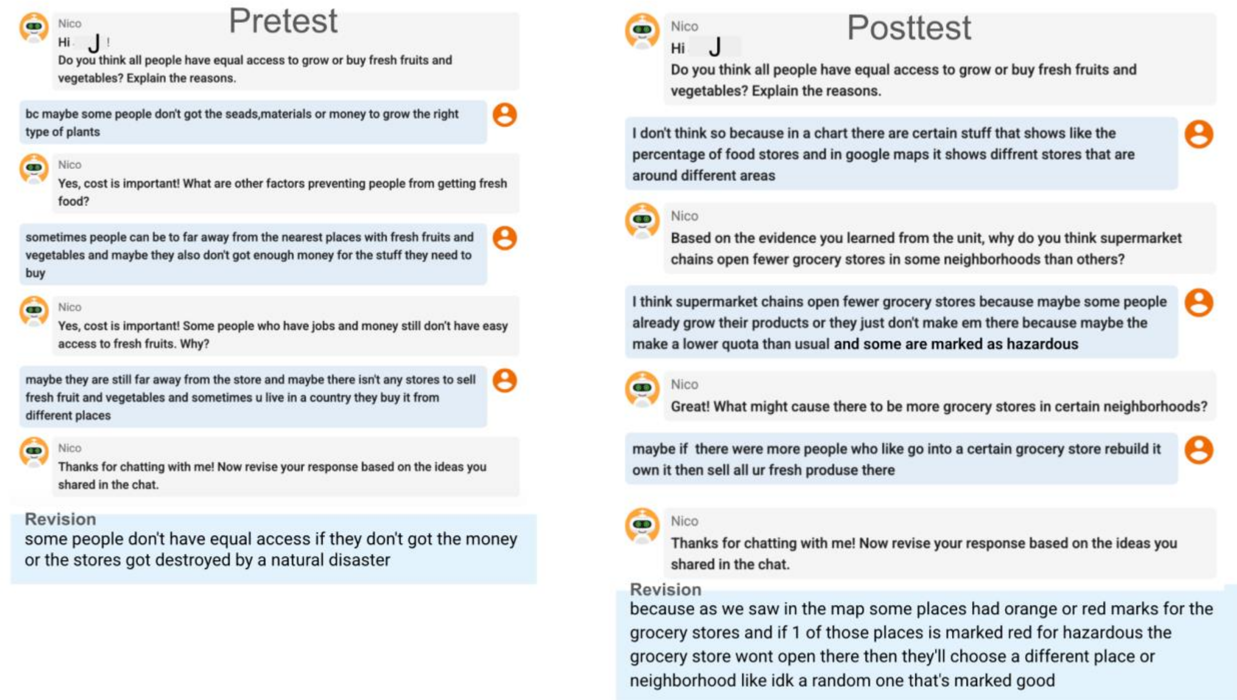

Figure 8. Student J's AI dialogs at pretest and posttest in Ben's class

Ben's case demonstrates that the dialog supports meaningful food justice reasoning when a teacher makes science accessible and personally proximate for students. Without extensive customization beyond the WISE unit, his 6th-grade students made consistent, significant KI gains through dialog interaction at both timepoints.

Students incorporated community-based evidence ideas that Ben's instruction had made personally observable: distance to grocery stores remained significant across both timepoints, reflecting the accessibility framing Ben sustained throughout the unit, and supermarket availability reached significance at posttest, reflecting the specific effect of his food access map discussion on students' ability to reason about the distribution of fresh food stores in their neighborhoods. Socioeconomic status was strongly elicited by the dialog at pretest, when it represented existing community knowledge students already held, but shifted to baseline knowledge by posttest. Idea 8, factors affecting plant photosynthesis and growth, showed modest and consistent but non-significant increases at both timepoints, suggesting the school garden and food access map activities introduced the science-to-food-justice connection but did not establish it as a central explanatory idea in students' reasoning. Structural ideas introduced through the WISE unit, including redlining and food apartheid, were present in instruction but showed no significant incorporation through dialog interaction. This pattern reveals the relationship between instructional emphasis and dialog incorporation: the dialog most reliably helps students elaborate ideas that instruction has made personally observable and nameable. Section 6 examines what becomes available when a teacher builds explicit customizations around both community-based and structural ideas through a problem-posing orientation.

6.2.2 Positioning Students as Transformative Intellectuals: Dialog and Instruction in Amy's Classroom

Amy's Pedagogical Orientation

Amy approached food justice instruction through what Morales-Doyle (2018) describes as a problem-posing, justice-centered orientation, positioning food justice not as a hook into science learning but treating relevance as political, relational, and co-constructed. This framing emerged from a relational understanding of her students built over two years of teaching many of them, and from a genuine belief in their capacity to observe, name, and act on injustice. She articulated her orientation experientially rather than theoretically: "When you teach issues of social justice, the way I see it is like you're giving students the language to what they're already observing." This was not abstract for Amy: she knew which students lived in food desert neighborhoods, which students recognized the local Mexican supermercado as their neighborhood grocery store, and which students had never had language for what they witnessed every day. Instruction, for Amy, was the process of connecting that existing observation to concepts, evidence, and action.

Amy's pedagogical goal came into focus when she reflected on her 8th graders' Aspiration Board projects after the implementation: "Every single one of their eighth grade projects really centered in this idea of changing food injustice, and we deserve access to healthy food here in [school]. And the HOW is kind of where they had some creativity." Student creativity and agency operated within the direction of understanding and transforming food insecurity.

Amy also attended carefully to the emotional dimensions of food justice instruction, navigating a tension that the KI+JCSP framework names but cannot fully operationalize: how to make structural inequity visible without overwhelming students whose own communities are the subject of study. She reflected on the video I chose to illustrate food apartheid, which was set outside her city: "I think it was helpful that it was not so local that it would bring up complicated emotions for students. It allowed us to talk about it in a different way." At the same time, she checked in individually with a student who lives in a food desert neighborhood, who told her: "I know, I live this." Amy reflected: "She was able to connect the injustice she already witnesses and now has a name for it." This attentiveness to emotional safety, ensuring food justice instruction empowered rather than overwhelmed, reflects a dimension of problem-posing pedagogy that Amy enacted through relationship and care rather than curriculum design alone.

Amy's Instructional Activities

Amy implemented the WISE unit over approximately 20 class periods across six weeks, designing local activities with me across multiple pre-implementation meetings. Where Ben's instruction enacted making SJSI accessible and making SJSI thinking visible through whole-class interaction, Amy enacted all four KI+JCSP extended principles across her activities, each addressing a distinct challenge of problem-posing pedagogy: connecting science concepts to students' lived community contexts, building structural vocabulary and critical consciousness, and facilitating authentic community action. Table 7 presents Amy's activities organized by KI+JCSP principle, ideas made available, and student learning observed.

Table 7. Amy's Customization Activities and Implementation Details

Activity	JCSP approach	KI+JCSP Principle	Instruction	Food Justice target ideas made available

Hands-on Garden Project (Weeks 3-4, Lessons 3-4)	Problem-posing, justice-centered	Making SJSI accessible at the community level: connecting science and school garden to students' own capacity to build food access	Students presented garden plans using scientific mechanism, voted democratically, planted seeds, observed growth, harvested; partnered with school gardening teacher	Availability of community gardens (Idea 6); factors affecting plant growth and photosynthesis, such as drought and shade tolerance (Idea 8)
Collaborative Food Access Map Discussion (Week 4, Lesson 5)		Making thinking visible + Helping students learn from others	Students worked in table groups to collect data on different numbers of grocery stores nearby; discussed findings with peers	Supermarket availability (Idea 5), Distance to grocery stores (Idea 9), SES (Idea 13), Food desert/apartheid (Idea 11)
Vocabulary Scaffolding: food desert vs. food apartheid (Week 5, Lesson 5)		Making SJSI accessible at the structural level: connecting SJSI vocabulary to students' existing observations	Built on 8th graders' prior knowledge of redlining from 7th grade; students distinguished between "desert" and "apartheid"	Food desert/apartheid (Idea 11), Systemic discrimination/redlining (Idea 14), Race/ethnicity injustice (Idea 18)
Aspiration Board Project (Week 5, Lesson 5)		Promoting autonomy: positioning students as transformative intellectuals designing and enacting real community interventions	Students worked in groups to chose own aspiration; genuine civic action including bilingual food bank posters, food justice speech to younger students	Social transformation (Idea 17), Food desert/apartheid (Idea 11), Systemic discrimination/redlining (Idea 14),

Hands-on Garden Project (Weeks 3-4, Lessons 3 & 4). Before the food justice lesson, Amy partnered with the school gardening teacher to extend the WISE garden challenge into a sustained hands-on project. Students worked in groups to design garden plans that balanced drought and shade tolerance, nutrition values, and costs. They presented their plans to the class, voted democratically for what to grow, planted seeds, observed growth, and eventually harvested. This activity enacted making SJSI accessible at the community level, grounding abstract photosynthesis and ecosystems in students' own school community and making the connection between science and food access materially real rather than illustrative. Amy reflected that having the project "circle back" to science concepts was what made the food justice connection feel coherent rather than tacked on: "I think that's why it was helpful to have the project kind of circle back to those [science concepts]." The garden itself became material evidence that communities can build food access, not just study its absence, making community gardens (Idea 6) and photosynthesis factors (Idea 8) personally observable before students encountered the food access map data.

Collaborative Food Access Map Discussion and Vocabulary Scaffolding (Week 4, Lesson 5). Amy brought seven extra laptops so students could explore the food access map in table groups rather than individually. Each table had a student assigned as data manager to record findings (Figure 9), 2-3 students assigned as data investigators to count and verify numbers, and 1-2 students assigned as data analysts to notice patterns. This structural choice enacted helping students learn from others: students developed observations through peer dialogue before sharing with the whole class, making the heterogeneity of students' community knowledge a resource rather than routing everything through the teacher. After independent table exploration, Amy facilitated a whole-class discussion: "Any patterns you noticed?" Students shared sophisticated observations: "J has many convenience stores but no supermarkets"; "X is in richer neighborhoods because they have more supermarkets that's more expensive." Rather than providing answers, Amy pressed students to connect their data to demographics and structural causes, making community-based evidence about supermarket availability, distance, and SES (Ideas 5, 9, 13) visible through students' own observations.

Click the graph to open the map. Count stores near each school area and fill in the table.

	School WO	School A	School CM
Total local grocery stores			
Total small convenience stores			
Total farmers' markets			
Total supermarkets			
Other food stores you find			

Figure 9. Group discussion with data analysis on food stores near school campus

Amy also attended to culturally relevant dimensions of the map data. When students raised CZ, a regional grocery chain carrying culturally meaningful foods, as a category the map did not capture, Amy navigated the moment carefully: "Yes, when we think of a grocery store near our school, we think of this place. And it is not as big as a supermarket but certainly serves great fresh produce." This move enacted the JCSP extension of making SJSI accessible: recognizing students' community knowledge as legitimate evidence rather than peripheral noise.

This discussion then created conditions for the food desert versus food apartheid distinction to emerge from students themselves, enacting making SJSI accessible at the structural level. Amy built on the foundation that her 8th graders had learned about redlining in 7th grade, deliberately creating a bridge between historical policy and current food access patterns. One student articulated: "Desert is like a natural geographic meaning, it takes away the intention behind it, like the race background." Another added: "Desert has no water at all, but food apartheid is not no food at all, it has food, but all junk food, not healthy fresh food." These student-generated distinctions made the vocabulary of structural critique accessible by connecting it to students' own observations of their neighborhoods, making food desert/apartheid (Idea 11), systemic discrimination/redlining (Idea 14), and race/ethnicity injustice (Idea 18) available in students' reasoning.

Aspiration Board Project (Week 5, Lesson 5). At the end of the food justice unit, Amy designed an Aspiration Board project (Figure 10) in which each table chose a food justice aspiration and created detailed action plans to address it, enacting promoting autonomy at the level of genuine civic action. The projects ranged widely in form but converged on substance. One group designed a bilingual food bank poster for the whole

school campus, calculated how many copies were needed at each building and entrance door, translated materials to Spanish, wrote an announcement Amy read at a teacher meeting, and visited all middle school classes to present. Another group created food justice slides targeting 3rd graders, reasoning that "third graders are responsible enough to bring home information, but they're not so young that they're not gonna understand," and presented during morning crew time in 3rd classrooms.

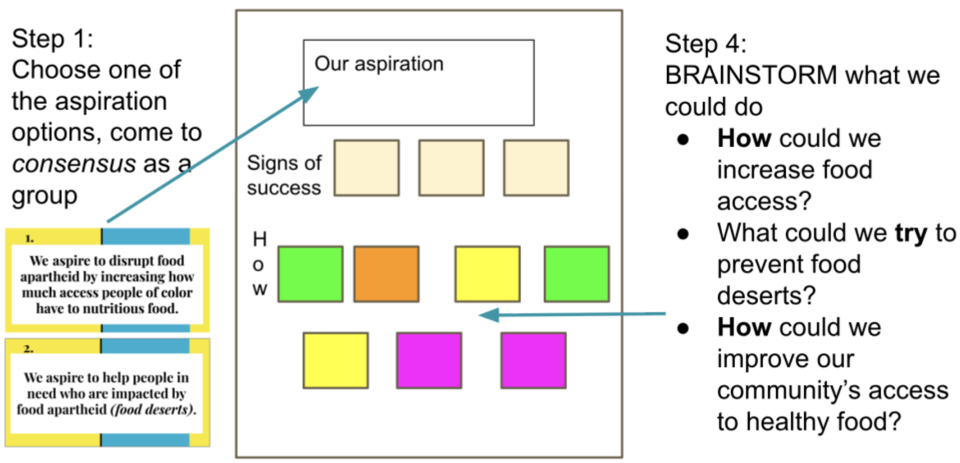

Figure 10. Aspiration Board Student Activity that Amy designed

Amy gave students genuine autonomy to act: "I wrote them a pass that gave them permission to be walking the halls and talking." She reflected that students she would not have expected to lead took ownership: "One group, a couple of the students in that group don't always make me feel super successful in class, they really got into the project and made this poster about the food bank to the whole campus including staff. They were so engaged." This positioning of students as community educators, not just learners, reflects Morales-Doyle's (2017) vision of students as transformative intellectuals, enacted through Amy's trust in her students rather than through curriculum design alone.

Amy's Student Learning Outcomes

The pattern of ideas students incorporated through dialog interaction reflects the cumulative effect of Amy's four customizations on students' knowledge repertoire. At pretest, before instruction, students showed the same starting point as Ben's students, incorporating primarily community-level evidence ideas through dialog interaction. At posttest, after Amy's hands-on garden project, collaborative food access map discussion, vocabulary scaffolding, and Aspiration Board project had built both community-based and structural ideas into students' knowledge repertoire, students incorporated both community-level and structural ideas, and the dialog helped them integrate those newly available ideas into linked explanations. The contrast between pretest and posttest idea incorporation is the quantitative results of Amy's problem-posing instruction.

KI Gains. Amy's students demonstrated significant KI gains through both dialog interaction and instruction. A CLMM revealed a significant main effect of dialog across both timepoints ($\beta = 2.45$, $SE = 0.50$, $z = 4.91$, $p < .001$), indicating that revised explanations were scored substantially higher than initial explanations. Gains were consistent: mean initial KI 2.26 (SD = 0.82) to mean revised KI 3.11 (SD = 0.86) at pretest (+0.94); mean initial KI 2.81 (SD = 0.90) to mean revised KI 3.67 (SD = 0.86) at posttest (+0.98). The main effect of instruction was also significant ($\beta = 1.51$, $SE = 0.48$, $z = 3.14$, $p = .002$), indicating that students' knowledge grew substantially across the unit through the cumulative impact of Amy's three classroom activities. The absence of a significant Dialog × Timepoint interaction indicates the dialog was equally effective before and after instruction. Table 8 presents raw mean KI scores.

Table 8. *KI Scores Across Timepoints in Amy's 8th grade Classrooms (N = 53)*

Timepoint	Initial KI M (SD)	Revised KI M (SD)	KI Gain
Pretest	2.26 (0.82)	3.11 (0.86)	+0.94
Posttest	2.81 (0.90)	3.67 (0.86)	+0.98

Figure 8 (right panel) presents the estimated marginal means from the CLMM. Amy's initial KI scores show a larger increase from pretest (2.24) to posttest (2.80) compared to Ben's, directly reflecting the significant main effect of instruction. Students arrived at the posttest dialog with richer, more structurally grounded knowledge to draw on, and the dialog helped them integrate it.

Ideas Added During Dialog. Wilcoxon signed-rank tests confirmed the dialog's scaffolding function at both timepoints. At pretest ($N = 45$), students added $M = 0.82$ new ideas ($SD = 0.96$; $V = 276$, $p < .001$), including $M = 0.53$ target ideas ($SD = 0.76$; $V = 153$, $p < .001$). At posttest ($N = 53$), students added $M = 1.30$ new ideas ($SD = 1.22$; $V = 630$, $p < .001$), including $M = 0.85$ target ideas ($SD = 0.89$; $V = 496$, $p < .001$). The marginally significant increase in ideas added from pretest to posttest (Total: $W = 932$, $p = .052$; Normative: $W = 948$, $p = .058$) suggests that Amy's instruction gave students a richer knowledge base to draw on during dialog interaction at posttest, consistent with the significant instruction effect in the CLMM.

Specific Ideas Incorporated. McNemar's tests with Benjamini-Hochberg correction identified which specific ideas students incorporated through dialog interaction (Table 9). The pattern show the cumulative effect of Amy's activities. At pretest, before instruction, students incorporated community-based evidence ideas: unequal access claim (+33.3%, $p < .001$), distance to grocery stores (+17.8%, $p = .005$), and socioeconomic status (+24.4%, $p < .001$). These are the same community-based evidence ideas Ben's students incorporated, reflecting a shared starting point before instruction. At posttest, students incorporated a wider range of ideas across both tiers. Among community-based evidence ideas, supermarket availability showed significant incorporation (+20.8%, $p = .002$), reflecting the collaborative map discussion in which students counted and compared the number of fresh food stores across neighborhoods. Among structural ideas, students incorporated redlining policy (+24.5%, $p = .002$) and food apartheid (+20%, p

= .003), reflecting the cumulative effect of Amy's vocabulary scaffolding and the peer dialogue that connected historical policy to students' own neighborhood observations. These structural ideas appeared at posttest because Amy's instruction had made them available in students' reasoning, and the dialog helped students integrate them into linked explanations.

Table 9. Ideas Added During Dialog (Initial to Revised Responses) in Amy's classrooms

Idea	Time	N	Initial %	Revised %	Change	χ^2	p (uncorr.)	p (BH)
1b: Claim - Not Equal Access	Pretest	45	13.30%	46.70%	33.30%	11.84	0.0006	0.0016
	Posttest	53	32.10%	66%	34%	16.2	0.0001	0.0004
8: Factors affecting Photosynthesis	Pretest	45	0.0%	9.4%	9.4%	1.33	0.2483	0.4964
	Posttest	53	4.4%	15.6%	11.2%	1.78	0.1824	0.7296
9: Distance to Grocery Stores	Pretest	45	2.20%	20%	17.80%	8	0.0047	0.005371
	Posttest	53	9.40%	37.70%	28.30%	15	0.0001	0.0004
13: Socioeconomic Status	Pretest	45	22.20%	46.70%	24.40%	8.07	0.0045	0.005371
	Posttest	53	13.20%	30.20%	17%	7.36	0.0067	0.0067
11: Food apartheid	Pretest	45						
	Posttest	53	7.30%	27.30%	20%	8.90	0.0028	0.0047
14: Redlining Policy	Pretest							
	Posttest	53	17%	41.50%	24.50%	9.94	0.0016	0.0032
7: Supermarket Availability	Pretest							
	Posttest	53	9.40%	30.20%	20.80%	9.31	0.0023	0.00368

Idea 8, factors affecting photosynthesis and plant growth, showed a consistent pattern across timepoints that reflects Amy's garden project in a specific way. At pretest, before instruction, no students mentioned plant growth in their initial responses, and 9.4% added it through dialog interaction. At posttest, after the hands-on garden project in which students designed plans, planted seeds, and observed growth, 4.4% mentioned plant growth in their initial responses and 15.6% in revised responses (+11.2%). The increase from pretest to posttest, while not reaching statistical significance ($p = .182$), reflects the cumulative effect of the garden project on students' readiness to connect science mechanisms of plant growth to food access when the dialog prompted them. Students who had actually grown plants were more prepared to use photosynthesis and growth conditions as evidence for unequal food access.

One idea central to JCSP goals, social transformation, was not reliably detectable by the model (F1 = 0.17) and does not appear in Table 6. But its absence from the quantitative results is itself evidence of something important: the ideas that matter most for justice-centered learning are precisely those that resist standardized detection. Qualitative evidence from observation fieldnotes and the Aspiration Board projects documents that social transformation ideas were present and active in Amy's classroom in ways the dialog could not capture. Students who visited elementary classrooms to present food justice materials, translated posters into Spanish, and wrote announcements read at teacher meetings were not just demonstrating knowledge; they were enacting their own capacity to change their community's relationship with food. The dialog created space for students to articulate what they knew; Amy's activities and relationships created the conditions for students to act on it. Quantitative analysis documents what the model can see; qualitative evidence documents what justice-centered instruction can reach beyond the model's ceiling.

Student S's dialog interactions (Figure 11) illustrate how the dialog helped students add and integrate ideas across both community-based evidence and structural policy, drawing on what Amy's instruction had made available. At pretest, the dialog prompted S beyond geographic observations about farms versus cities, eliciting distance and store availability as barriers, but S's revision remained at the community-based evidence level. At posttest, the dialog's guidance produced substantially richer responses because Amy's instruction had built a wider range of ideas into S's knowledge repertoire. When the dialog asked why some neighborhoods have fewer grocery stores, S drew on Amy's vocabulary scaffolding to add redlining explicitly as a structural cause, an idea the pretest dialog had not been able to elicit because Amy's instruction had not yet occurred. When the dialog prompted S to think about action, S proposed spreading food justice information and collecting fresh fruit from a garden to give to those in need, drawing on the social transformation reasoning Amy's Aspiration Board project had made available. S's posttest revision integrated community-based evidence, structural policy, plant science mechanisms, and community action into a single coherent explanation. The dialog provided the consistent scaffolding that prompted S to articulate and link these ideas; Amy's instruction determined what ideas S had to draw on.

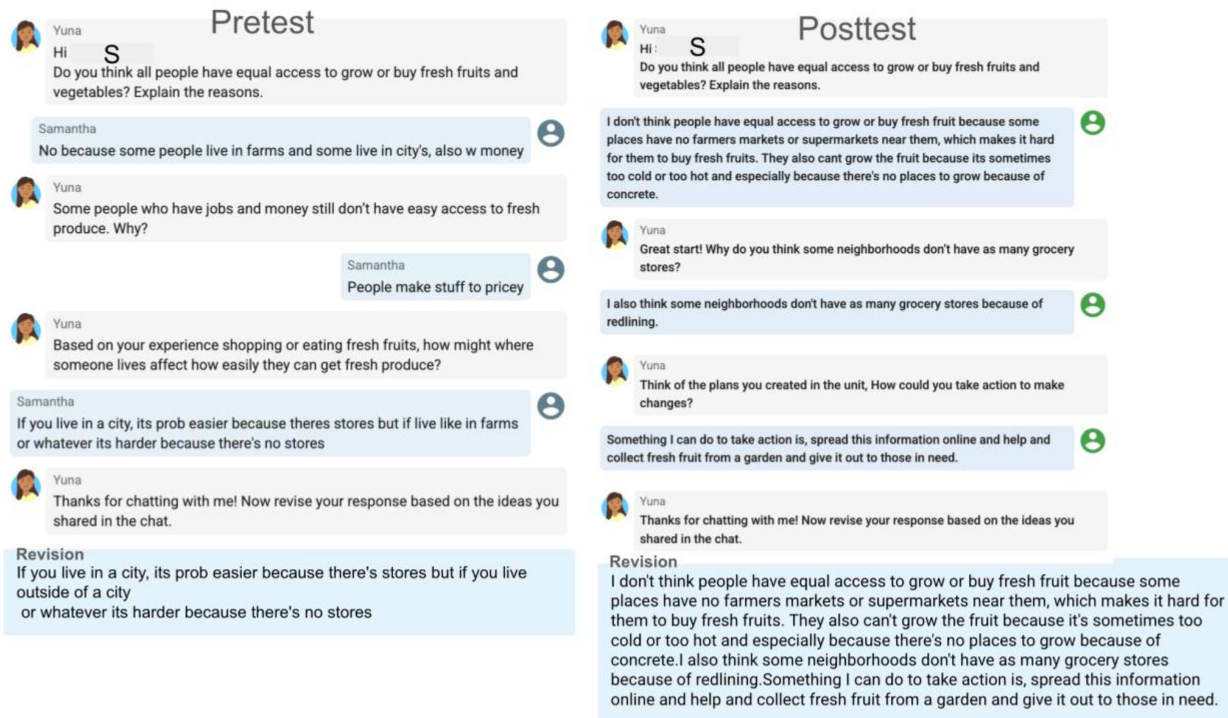

Figure 11. Student S's AI dialogs at pretest and posttest in Amy's class

Both teachers' customized dialogs and classroom practices produced significant knowledge integration gains, with the dialog generating robust within-session improvements in both classrooms (Amy: $\beta = 1.63$, $p < .001$; Ben: $\beta = 1.75$, $p < .001$) across all four timepoints. Across the instructional unit, Amy's students showed significant KI gains from pretest through posttest, with the distribution shifting from 64% at KI 3 and 36% at KI 2 at pretest to 20% reaching KI 4, 68% at KI 3, and only 12% at KI 2 by posttest. The KI gain was still significant in delayed posttest. In post-implementation interviews Amy noted: "I saw students connecting what they learned back to the previous ecology lesson we did earlier this year," reflecting her observation that students were building on prior knowledge rather than accumulating isolated facts. Ben's students showed significant KI gains from pretest through posttest, with the distribution shifting from 60% at KI 2 and 35% at KI 3 at pretest to 37% at KI 2 and 60% at KI 3 by posttest. Ben observed: "Some of them could tell me what goes in and what comes out, but connecting that to the bigger energy transfer story...that's still developing," reflecting his awareness of where students were in their learning progression and his vision for the next instructional steps. This noticing positions Ben to build on this foundation in future units. In both cases, classroom practices and dialogs functioned synergistically: activities prepared students to engage more productively with dialog prompts at later timepoints, and the specific ideas that developed differed based on each teacher's pedagogical emphasis. Full KI score and idea frequency data are presented in Appendix Table 6.

7 Cross-Case Comparison: Pedagogical Orientation and Dialog Effectiveness

Ben and Amy implemented the same food justice dialog and WISE unit with different pedagogical goals and levels of teacher customization. Ben's approach centered on making food justice personally accessible and proximate for 6th graders through teacher-mediated whole-class interaction, what he himself characterized, drawing on Morales-Doyle (2018), as an “injected relevance” orientation with one teacher-designed activity beyond the WISE content over four weeks. Amy's orientation centered on positioning 8th graders as transformative intellectuals through four classroom activities that built peer dialogue, structural vocabulary, and genuine community action over six weeks. Both implementations reflected instructional decisions grounded in each teacher's knowledge of their students, their grade levels, and their theoretical orientation toward food justice instruction. Table 10 summarizes the key dimensions of their implementations through the KI+JCSP analytical lens.

Table 10. Teacher instructions and customizations during the unit

Dimension	Ben	Amy
JCSP Approach	food justice as motivational context for science learning	food justice as core learning objective and site of community action
KI+JCSP Principles Enacted	Making SJSI accessible (personal proximity); Making SJSI thinking visible (teacher-mediated synthesis)	Making SJSI accessible (structural vocabulary distinction); Making SJSI thinking visible (student-generated data report); Helping students learn from others (peer discussion); Promoting autonomy (aspiration board projects)
Time Investment	~12 class periods (4 weeks) • Focused on core WISE content • Limited customization	~20 class periods (6 weeks) • Added group projects, peer feedback • Created custom templates and worksheets
Food Access Map Activity Instruction	Whole-class guided discovery • Corrected misconceptions ("What's a convenience store?") • Teacher-led pattern identification (“How many yellow dots does [school] have?”	Table-based collaborative exploration + Whole-class discussion • Student-developed data strategies • Student-led pattern presentation • Teacher addressed confusion if needed

Action Focus	Proximal and personal: school garden, food bank resources, "help people immediately near to us"	Authentic community action: bilingual posters, food justice presentations, community food bank awareness
Ideas Students Integrated	Community-based evidence: distance, SES, unequal access	Community-based evidence + systemic injustice: redlining, food apartheid, supermarket availability, social transformation

Three findings emerge from comparing the two cases. First, the dialog's effectiveness was robust across large-scale benchmark assessments and in two very different instructional contexts. Ben and Amy differed in grade level, pedagogical orientation, customization level, interaction structure, and time investment, yet both classrooms showed significant and consistent KI gains through dialog interaction at both pretest and posttest. This consistency demonstrates that two iterations of RPP design produced a dialog with sufficient breadth and accuracy to scaffold food justice reasoning across diverse classroom contexts, providing a reliable baseline regardless of how deeply teachers customized instruction around it.

Second, teachers' pedagogical orientation shaped which ideas students incorporated through dialog interaction. Both teachers' students began at a similar place: at pretest, before any instruction, students in both classrooms incorporated community-based evidence ideas rooted in everyday observation, including distance to grocery stores, socioeconomic status, and unequal access claims. After Ben's whole-class food access map discussion, in which he guided students to distinguish the number and type of fresh food stores across neighborhoods, students expressed significantly more ideas on supermarket availability at posttest, showing that teacher-mediated whole-class activities make community-level patterns visible and nameable. Ideas that require explicit vocabulary work, such as redlining and food apartheid, were present in Ben's instruction through the WISE unit but were not the focus of his customizations, and did not incorporate significantly into students' reasoning. Amy's justice-centered orientation extended students' reasoning further: her collaborative map discussion produced significant supermarket availability incorporation through student-generated peer data, her vocabulary scaffolding made redlining and food apartheid available as ideas students could name and use as evidence, and her Aspiration Board project positioned students as transformative intellectuals acting on food injustice in their community. The dialog revealed these differences: the same dialog, running as a consistent probe before and after instruction, captured the distinct imprint of each teacher's pedagogical orientation on students' reasoning.

Idea 8, factors affecting plant photosynthesis and growth, tells a complementary story about the science-to-food-justice connection the unit was designed to build. Neither teacher's instruction drove this idea to statistical significance in the dialog, but the pattern differed in ways that reflect each teacher's approach. Ben's students showed modest and consistent dialog-elicited increases at both timepoints, suggesting the food access map observations and the emphasis on school garden introduced the connection between science and food access, but it did not become a central explanatory idea in students' reasoning without sustained hands-on science experience. Amy's students showed a larger increase at posttest (+11.2%) compared to pretest (+9.4%), reflecting the cumulative effect of the hands-on garden project: students who had actually designed garden plans, planted seeds, and observed growth were more prepared to draw on photosynthesis and plant growth conditions as actions for improving unequal food access when the dialog prompted them. Student work from Amy's classroom shows that students explicitly reasoned about drought tolerance and shade conditions as constraints on what communities can grow in their garden designs, and in their Aspiration Board projects, one group connected plant biology directly to food justice action, reasoning that drought conditions drive up the cost of fresh fruits and vegetables and designing a campaign to educate their community about more affordable fresh food alternatives.

Third, both teachers designed instructional activities outside the dialog that built the ideas and connections students drew on when they engaged with it. Ben's whole-class discussions and food access map synthesis helped students name and link community-based evidence that the dialog then helped them articulate and elaborate. His students also made personal connections the dialog had not anticipated, such as independently exploring food bank resources on WISE for their own families, reflecting the proximal and personally relevant framing Ben sustained throughout the unit. Amy's peer dialogue, vocabulary scaffolding, and Aspiration Board project built ideas about systemic injustice and community action that the dialog then helped students integrate into linked explanations. In Amy's case, some of what students learned through these activities, particularly the food desert versus food apartheid distinction and the social transformation reasoning in the Aspiration Board projects, reflected reasoning the model could not reliably detect and therefore could not scaffold on its own. Together these cases show that teacher instruction and dialog scaffolding are complementary: the dialog extends the reach of instruction for community-based evidence ideas it can reliably detect and respond to, and teacher-designed activities build the ideas and connections that determine what students bring to the dialog in the first place.

8 Discussion

Through two iterations involving 798 and 869 students, the RPP developed a dialog reliably detecting 22 food justice ideas (micro F1 = 0.73), organized into community-based evidence and systemic injustice ideas. Both focal teachers' classes showed significant dialog-driven KI gains (Ben: $\beta = 2.48$, $p < .001$; Amy: $\beta = 2.45$, $p < .001$). However, the pattern of idea incorporation differed systematically. Ben made food justice personally accessible and proximate for his 6th graders; his students incorporated community-based evidence ideas at both timepoints, extending to supermarket availability at posttest. Amy positioned her 8th-grade students as transformative intellectuals through a problem-posing orientation; her students incorporated both community-based evidence and structural ideas at posttest, including redlining (+24.5%, $p = .002$) and food apartheid (+20%, $p = .003$), and her instruction produced a significant main effect beyond dialog gains ($\beta = 1.51$, $p = .002$).

8.1 RPP Design Builds Effective Dialogs for Food Justice Reasoning

The two-iteration design process produced a dialog that achieved meaningful model accuracy (F1 = 0.74 after Iteration 1; F1 = 0.73 after Iteration 2), a rubric that captured the full range of student reasoning from community-level evidence to structural critique, and significant learning gains across two classrooms. This demonstrates that the RPP design framework developed for traditional science topics (Chapter 2) can be adapted to social justice science issues, despite the additional complexity of detecting ideas that integrate scientific concepts with lived experience and structural inequity.

Two features of the design process were particularly important for the food justice context. First, the dual-criteria approach of evaluating ideas by both model performance and instructional value was necessary. For food justice, students tend to express ideas in diverse ways based on their lived experiences, such as social transformation, culturally relevant food, race/ethnicity correlation, which are the ideas technically difficult to detect. Without an explicit criterion for instructional value that could override technical feasibility, the design process would have systematically removed the ideas that matter most for justice-centered learning. This finding extends prior RPP design work (Gerard, Bradford, Debarger, et al., 2022; Penuel et al., 2011) by showing that partnership structures must explicitly protect pedagogical priorities against technical optimization pressures, particularly when the learning domain involves equity and justice.

Second, authentic classroom data was necessary for capturing the full range of student reasoning about food justice. The 798 benchmark responses used to build the Iteration 1 model did not reveal the multi-turn reasoning progressions that emerged when 869 students engaged with the dialog in classroom contexts. New distinctions between systemic discrimination, vague discrimination, and observed race/ethnicity correlation only became visible when students reasoned through dialog sequences embedded in instruction. This confirms McKenney and Reeves' (2018) argument that design-based research requires iterative cycles grounded in authentic practice, and extends it to the specific case of AI dialog development: benchmark data alone is insufficient for capturing the reasoning patterns that matter most in classroom contexts.

8.2 Dialog Scaffolds Accessible Community Evidence

Ben's case demonstrates that the dialog supports meaningful food justice learning when enacted through a relevance and accessibility orientation. Significant KI gains through dialog interaction at both pretest and posttest show that a teacher who makes food justice personally accessible and proximate for students, through teacher-mediated whole-class discussion and one added activity, can expect robust learning gains from the dialog. This finding has practical significance for scaling justice-centered science instruction: not every teacher needs to enact a full problem-posing orientation to benefit from a well-designed AI dialog.

This finding extends the KI framework's account of effective science learning (Linn & Eylon, 2011) to a social justice science context. The KI principles of making SJSI accessible and making thinking visible, which characterized Ben's instructional approach, produced significant dialog gains when grounded in a dialog built from authentic student responses and designed with JCSP-aligned guidance. Prior work combining KI and JCSP frameworks has focused primarily on teacher-designed instruction (Bradford, Gerard, et al., 2023); this chapter shows that AI dialogs grounded in both frameworks can extend JCSP-aligned scaffolding to every student who engages with the dialog, regardless of the teacher's pedagogical orientation.

At the same time, Ben's marginal instruction effect suggests that exposure to structural ideas including redlining and food apartheid through the WISE unit, without deeper instructional elaboration through peer dialogue or explicit vocabulary work, may not be sufficient to bring these ideas into students' active reasoning. This is consistent with Morales-Doyle's (2017) argument that justice-centered science pedagogy requires more than content exposure: it requires creating conditions for students to connect structural inequity to their own observations and experiences.

8.3 Dialog Integrates Structural Reasoning

Amy's case demonstrates that a problem-posing orientation, enacted through three teacher-designed local activities, amplifies the dialog toward JCSP goals the dialog alone cannot achieve. The significant main effect of instruction in Amy's classroom reflects the cumulative impact of her activities on students' knowledge repertoire. The mechanism is specific: Amy's vocabulary scaffolding and collaborative map discussion built structural knowledge about redlining, food apartheid, and supermarket availability through peer dialogue and political vocabulary work, and the dialog then helped students integrate those ideas into linked explanations. Without those activities, the dialog had no structural knowledge to scaffold. Students arrived at the posttest dialog with a richer, more structurally grounded knowledge base, and the dialog helped them consolidate and articulate what instruction had made available.

Amy's instructional approach also illustrates what justice-centered instruction requires beyond curriculum design. Her attentiveness to emotional safety, her knowledge of individual students' neighborhood contexts, and her two years of relationship with these students made the Aspiration Board projects possible in ways the WISE unit alone could not have anticipated. The food desert versus food apartheid distinction that emerged from Amy's collaborative map discussion was a product of peer dialogue that the dialog created space for but could not produce alone. The community action projects that emerged from the Aspiration Board positioned students as transformative intellectuals in ways that the model could not detect but that qualitative evidence documents clearly. This suggests that AI dialogs designed for JCSP contexts should be evaluated not only by what gains they produce independently, but by how effectively they amplify what teachers with a problem-posing orientation can accomplish.

This finding extends the KI framework's account of teacher noticing and relational knowledge in science learning (Luna, 2018; Wilkerson et al., 2016) to the SJSI context. Amy's instruction took students' community knowledge seriously as legitimate evidence, built structural vocabulary through student-generated discourse, and positioned students as capable of sophisticated reasoning and community action. These moves enacted the full range of KI+JCSP extended principles: making SJSI accessible at the structural and political level, making thinking visible through student-generated synthesis, helping students learn from others through peer dialogue, and promoting autonomy through genuine civic action. The dialog provided consistent scaffolding at scale across every student who engaged with it; Amy's relationships and pedagogical judgment created the conditions that determined what students brought to that scaffolding.

8.4 Teacher Pedagogical Orientation and AI Scaffolding Play Complementary Roles

Together, Ben's and Amy's cases reveal a complementary relationship between teacher instruction and AI scaffolding that extends prior work on technology-enhanced science learning (Linn & Eylon, 2011) in an important direction. In justice-centered science topics, the relationship between instruction and technology scaffolding is complementary: teacher instruction does not just add to what the dialog does, it shapes what the dialog can do by determining which ideas are available in students' reasoning when they engage with it.

Running as a consistent probe before and after instruction in both classrooms, the dialog revealed the distinct imprint of each teacher's pedagogical orientation on students' reasoning. Ben's relevance and accessibility framing made community-based evidence ideas personally observable and proximate, and the dialog helped students elaborate those ideas into linked explanations. Amy's problem-posing framing made structural ideas available through peer dialogue and political vocabulary, and the dialog helped students integrate those newly available ideas at posttest. The dialog did not create these differences. It detected them. The same rubric, designed through iterative RPP design to recognize the full range of food justice ideas from community observation to structural critique, was sensitive enough to capture what each teacher's orientation made possible.

This distinction has design implications: AI dialogs for SJSI should be designed not as standalone learning tools but as components of instructional systems where teacher activities and dialog scaffolding are deliberately sequenced to build on each other. It also has evaluation implications: the effectiveness of an AI dialog for SJSI cannot be assessed independent of the pedagogical context in which it is embedded, because that context shapes what the dialog can reveal.

8.5 Model Detection Limits Reveal Where Teacher Instruction Matters Most

The persistent difficulty detecting race, discrimination, and social transformation ideas reveals something important about the nature of justice-centered reasoning with direct implications for how AI systems should be designed and evaluated for SJSI contexts. Students' expressions of structural critique are linguistically diverse, contextually embedded, and emotionally situated in ways that resist standardized detection. As Bang and Medin (2010) argue, different communities bring different ways of knowing to science, and these ways of knowing are expressed in language patterns that dominant assessment frameworks systematically fail to recognize. The food justice dialog's low F1 scores for social transformation and race/ethnicity ideas reflect this broader challenge: the ideas that matter most for equity-oriented learning are precisely those that are hardest to standardize.

This finding connects to Philip et al.'s (2017) argument that equity-oriented STEM education requires attending to the racialized and political dimensions of learning that conventional assessment frameworks render invisible. The model's inability to reliably detect social transformation reasoning does not mean students are not reasoning about transformation. Amy's Aspiration Board projects demonstrate that they are. It means the model cannot see it. Designing AI systems for SJSI requires being explicit about what models cannot see, and designing instructional systems that ensure teachers can see it instead.

Practically, this suggests that AI dialogs for justice-centered science topics should be accompanied by teacher-facing analytics that identify where students' structural reasoning is developing, directing teacher attention toward the moments and students where human judgment and relational knowledge matter most. The model's limitations, made transparent through the RPP design process, became shared knowledge among teachers, researchers, and computer scientists that shaped how teachers planned instruction. That transparency is itself a design achievement worth replicating.

9 Limitations and Future Directions

Several limitations constrain the conclusions of this chapter. First, both focal teachers are long-standing RPP members with deep familiarity with KI frameworks and design processes. Ben explicitly connected his pedagogical approach to Morales-Doyle's framework through his master's program coursework; Amy brought years of social justice teaching experience and RPP collaboration to her implementation. Their implementations may not reflect what teachers with less RPP experience or theoretical background would do with the same dialog and unit. Future research should examine how teachers with varying levels of RPP involvement and familiarity with JCSP frameworks implement the food justice unit, and what support structures help less experienced implementers achieve comparable learning gains.

Second, the comparison between Ben and Amy conflates multiple dimensions simultaneously: grade level (6th vs. 8th), dialog version (Iteration 1 vs. Iteration 2), pedagogical orientation, and prior student knowledge of redlining from 7th grade instruction. It is not possible to isolate which of these factors accounts for the differences in instruction effects and structural idea incorporation. The cross-case comparison is interpreted as illustrative of what different pedagogical orientations make possible with the same shared dialog, not as a controlled comparison of customization depth. Future research should design studies that hold more of these dimensions constant, for example comparing two teachers at the same grade level implementing the same dialog version with different pedagogical orientations toward the same SJSI topic.

Third, the model's persistent difficulty detecting race, discrimination, and social transformation ideas limits the dialog's ability to scaffold the JCSP goals most central to justice-centered science learning. While this limitation points to teacher roles rather than design failure, it also means that students whose structural reasoning is most sophisticated may receive the least responsive adaptive guidance when the dialog cannot recognize and build on their highest-level ideas. Addressing this limitation likely requires approaches beyond training data expansion, including more flexible detection methods that can recognize diverse linguistic expressions of structural critique and culturally embedded ways of reasoning about inequity. Our evaluation relied on automated metrics (e.g., F1) and KI scores; future work should incorporate fairness and intersectionality-aware evaluation (Mangal & Pardos, 2024) along with richer forms of analysis, such as longitudinal tracking of idea development, or mixed-method studies that examine how AI-mediated feedback unfolds during instruction. Although teacher input and iterative refinement reduced some risks of bias, the model may under-represent students who use less common linguistic forms; bringing students into the RPP as co-designers could surface overlooked perspectives.

Fourth, this chapter examines a single SJSI topic, food justice, in a single geographic context with a student population that has direct lived experience of food access inequity. The dialog and unit were designed with this context in mind; their effectiveness in different geographic, demographic, or SJSI topic contexts is unknown. Future research should examine how the RPP design framework applies to other social justice science issues and whether dialogs built for one justice-centered topic transfer to others or require topic-specific design from scratch.

Chapter 5: Discussions and Conclusion

This dissertation set out to understand how AI adaptive dialogs can be designed, iteratively refined, and customized in partnership with teachers, to support middle school students' integration of complex science and social justice ideas in culturally and linguistically diverse classrooms. Three empirical chapters, spanning four years of collaborative design research across six diverse schools and more than 5,000 student responses, build a cumulative answer to that question. This concluding chapter synthesizes what those three chapters reveal when read together, identifies cross-cutting themes that are only visible in synthesis, and draws out implications for AI dialog design, science education practice, and future research.

1 Summary of Findings

Chapter 2 established that identifying and responding to the full range of student ideas in AI dialogs requires iterative refinement, and that teachers are essential contributors to that process. The most consequential driver of improvement across four design iterations was teacher-informed rubric expansion: when RPP teachers identified ideas that students frequently expressed but the initial rubric did not capture, incorporating those ideas improved both the NLP model's detection accuracy and students' knowledge integration gains. The chapter also showed that guidance type matters: distinguishing prompts outperformed eliciting prompts for students with detectable prior ideas, while eliciting prompts served students with no detectable ideas. These findings established the proof of concept for teacher-informed iterative design at the collective co-design level.

Chapter 3 revealed that teacher pedagogical goals shape how teachers enact the dialog in their classrooms. The RPP's design activities and structures, including workshops, pre- and post-implementation meetings, and classroom observations, supported Amy and Ben to articulate and refine distinct pedagogical goals that organized their customizations across multiple levels simultaneously: dialog guidance revisions, classroom activity design, and in-the-moment instructional moves. Amy's goal of making ecosystem connections visible and Ben's goal of making science accessible produced different but complementary learning outcomes, both statistically significant. This showed that teacher-informed customization makes AI dialogs more responsive to specific students in specific contexts.

Chapter 4 extended this work to food justice, a social justice science issue (SJSI) where student ideas are shaped by lived experience of structural inequality. The chapter showed that building an effective AI dialog for SJSI topics requires first building shared understanding within the RPP of what integrated SJSI reasoning looks like in students' own words. Food access reasoning is complex, drawing on personal experiences, community observations, and systemic critique. The chapter also showed that teachers' pedagogical orientations toward social justice teaching shaped not just how much students learned but which ideas became available in their reasoning. And it showed that the systemic injustice ideas most central to justice-centered science learning, such as redlining policy, racial discrimination, and social transformation, were the ideas the model found most difficult to detect reliably, pointing toward the essential role of teacher instruction in justice-centered learning contexts.

2 Discussion

2.1 Idea Detection Augments KI Scoring

A foundational contribution of this dissertation shows that combining idea detection with KI scoring gives teachers and researchers a rich picture of student thinking. KI scoring captures how well a student links ideas into a coherent explanation, and is the primary measure of learning in all three chapters. It answers the question: how integrated is this student's reasoning overall? But a holistic integration score, on its own, does not tell a teacher which specific ideas a student holds, which are missing, and which are expressed vaguely. A student scoring KI 3 at the food justice AI dialog might be missing socioeconomic evidence entirely, or might have distance to grocery stores but lack any structural reasoning about why that distance exists. The KI score does not distinguish between these cases.

Idea detection fills that gap by identifying which specific ideas are present in a student's response through dialog interaction. The McNemar and GEE analyses in Chapters 3 and 4 showed not just that students made KI gains, but which ideas moved significantly through dialog interaction and which did not, and how that pattern differed across teachers and timepoints. For example, in Chapter 3, both teachers' students significantly improved their explanations at the posttest, but the key ideas added were different. For Amy, her students added more ideas related to energy transfer and transformation in ecosystems, such as cellular respiration. For Ben, his students added more ideas related to photosynthesis chemical reactions.

Together, KI scoring and idea detection enrich the picture we could offer to teachers about students thinking. KI scoring shows whether students are integrating their ideas into linked explanations. Idea detection shows what those ideas are. For teachers, this combination is directly actionable: knowing that most students scored KI 3 tells you they are making partial link explanations; knowing that the most common ideas are animals directly using sunlight to keep warm tells you what instruction needs to make available next. The RPP dashboard (Wiley, et. al., 2024) translates this combination into a form teachers can use in real time, surfacing both aggregate KI patterns and idea frequency counts organized by rubric category. This combination of holistic and granular measurement is a design guideline worth carrying into future AI dialog development for science education.

2.2 Collective RPP Co-Design and Local Customization Form an Iterative Loop

The central argument of this dissertation is that teacher expertise within RPP operates at two distinct levels: collective RPP co-design, through which teachers shape AI dialogs and units that work broadly across diverse classrooms, and local customization, through which individual teachers amplify what those dialogs can accomplish for specific students in specific community contexts. What the three chapters together reveal is that these two levels form an iterative loop in which each level continuously informs and improves the other.

The entry-point idea expansion in the Photosynthesis dialog (Chapter 2) illustrates this iterative refinement most clearly. In RPP review meetings, teachers and researchers collectively identified that students frequently expressed broad, vague ideas as starting points: ideas like "the sun helps plants survive" that were scientifically incomplete but represented genuine prior knowledge worth building on. The RPP incorporated these into the expanded rubric, and the guidance designed around them worked well across classrooms: many students receiving guidance on this idea added more ideas. Teachers customized the guidance differently in their classrooms. Amy, pursuing her goal of making ecosystem connections visible, used the dialog's responses to entry-point ideas as a launching pad for deeper mechanistic reasoning to encourage students to trace how energy moved through the organisms they had named. Ben, pursuing his goal of linguistic accessibility, used the same responses to help emergent bilingual students build confidence and add new ideas in accessible language. Their different uses revealed different possibilities in the same guidance. When those observations fed back into RPP workshops, the team discussed which revisions to adopt collectively that then benefited all partner classrooms, not just Amy's and Ben's.

This loop is what makes teacher-informed design generative rather than merely responsive. Local customization does not just adapt a fixed tool to a local context. It generates knowledge about what the tool can and cannot do, which feeds back into collective improvement, which enables richer local customization, which generates further knowledge. Understanding this loop has implications for how RPPs should be structured: the value of sustained, multi-year partnerships lies not just in accumulated trust but in the accumulated design knowledge that this iterative loop produces over time.

2.3 Strengths and Limitations of AI Guidance

Across the three chapters, the AI dialogs demonstrated consistent strengths and persistent limitations that together clarify what AI guidance can and cannot do in justice-centered science learning contexts.

The dialog's strengths were most visible in its ability to accurately detect and respond to a wide range of student ideas in real time across diverse classrooms. In Chapter 2, the photosynthesis dialog reliably detected both target science ideas and vague entry-point ideas, and delivered guidance that helped students add new ideas and improve KI scores significantly across four iterations (Figure 8, Chapter 2). In Chapter 4, the food justice dialog achieved an overall F1 of 0.73 across 19 detectable ideas (Table 4, Chapter 4). Both dialogs produced significant KI gains through dialog interaction alone, even in benchmark conditions without instruction.

Across the three chapters, some ideas proved consistently harder for the NLP models to detect reliably. This is because of how students expressed them. Ideas that are linguistically diverse, contextually embedded, or emotionally situated are difficult for NLP models to capture because students express them through personal experiences, culturally specific examples, and language that varies widely across individuals and communities. In the food justice context, ideas like culturally relevant food, social transformation, and race and ethnicity injustice had among the lowest F1 scores in the rubric. This persistent challenge is worth naming explicitly.

Teacher-informed AI dialog design scales up responsive teaching by having the AI do reliably what it can do: detect a wide range of student ideas in real time, deliver personalized guidance based on those ideas, and surface patterns across a whole classroom that no single teacher could track manually. This creates space for teachers to focus on what requires their direct engagement. The ideas that are hardest to detect are often the ideas that most require a teacher’s attention: leveraging the diversity of students' community knowledge as a collective resource, holding safe space for feelings of systematic inequity, fostering students' reflections on their potential biases, developing disciplinary understanding to make sense of and address local social justice issues. Those are the irreplaceable core practices of justice-centered teaching.

The RPP co-design process documented in this dissertation shows how to navigate these strengths and limitations responsibly. In both the photosynthesis and food justice dialogs, teachers and researchers together evaluated which ideas to include in the rubric based on two criteria: model performance and instructional value. This dual-criteria approach made visible the ideas the model could not yet detect as teaching targets. Teachers already knew from classroom experience which ideas students struggled to articulate, and the dialog's detection data confirmed and extended those observations at scale, surfacing patterns across entire cohorts that no single teacher could track manually. The dialog augmented teacher noticing. For the ideas the rubric retained, classroom activities augmented the dialog by building on the gaps and patterns it identified. Ben's whole-class food access map synthesis made supermarket availability and distance to grocery stores personally observable and nameable. Amy's vocabulary scaffolding built on the dialog's detection gap for food apartheid and redlining by making those ideas available through peer dialogue, and the Aspiration Board project created conditions for social transformation reasoning that the model struggled to detect but that Amy had long observed in her students' commitments to their community. In each case, the dialog identified where students' thinking needed support; teacher-designed activities built the knowledge that the dialog could then help students articulate and integrate.

2.4 From Science Accessible to Justice-Centered: KI and JCSP Goals in an Integrated Ecosystem

A productive tension runs through Chapter 4: the photosynthesis dialog and the food justice dialog are embedded in the same WISE school garden unit, but they address different ideas and the connection between them was present but modest in student dialog data. Only one idea, factors affecting photosynthesis and plant growth in the food justice dialog, bridges the two dialogs directly, and student idea changes for this idea were not statistically significant. This is worth examining as a window into what makes this dissertation's approach distinctive and into how KI and JCSP goals operate together in an integrated instructional system..

The photosynthesis dialog does not ask students to memorize chemical reactions. Drawing on KI principles of making science accessible and eliciting the full range of student ideas, it connects photosynthesis to students' everyday understanding of how animals use energy from the sun to survive. It detects not only target science mechanisms but also vague and intuitive ideas. Students who write that "the sun keeps animals warm" or that "animals eat plants to get their everyday nutrition" are expressing genuine prior knowledge that the dialog recognizes and builds on rather than penalizes (Linn & Eylon, 2011). These entry-point ideas (see Figure 6 in Chapter 2) are the starting material for KI's process of distinguishing and linking: the dialog guides students from intuitive ideas toward mechanistic reasoning about energy flow through organisms. This is what makes the photosynthesis dialog go beyond typical photosynthesis instruction: it honors students' existing ideas about the environment around them and uses them as a foundation for deeper scientific understanding, grounded in the school garden phenomenon that students can directly observe and tend.

The food justice dialog then extends this foundation by pushing students to ask another KI question: Does this science work the same way for everyone? Why? Students who understand that plants need sunlight, water, carbon dioxide, and specific soil conditions to grow can ask why some communities cannot access the fresh produce those plants produce. The photosynthesis dialog establishes the science; the food justice dialog uses it as evidence for reasoning about structural inequity. Together, the two dialogs enact what the integrated KI+JCSP framework calls for: rigorous disciplinary learning that develops students' capacity to use science to understand and address injustice in their own communities.

There are many food justice curricula that help students reason about structural inequity, and many science curricula that help students understand photosynthesis and ecosystems. What distinguishes this dissertation's approach is that both are anchored in the same school garden phenomenon. Students design garden plans using science knowledge, weighing drought tolerance, shade conditions, nutrition values, and costs. That same garden then becomes the entry point for reasoning about why some communities can grow or access fresh food and others cannot, and for designing action to change it. Students who have actually designed garden plans, planted seeds, and harvested produce are not just learning about food access, they are building it in their own school community as transformative intellectuals (Morales-Doyle, 2017). The school garden is not just a motivational hook. It is the integrating context that makes the science and the justice reasoning feel like parts of the same inquiry, and that positions students as agents within that inquiry rather than observers of it.

The open question this framing leaves for the field is how AI tools can be designed to better support all three JCSP goals within a single coherent interaction by recognizing and responding to students' expressions of community identity, cultural knowledge, and commitment to change, without reducing those expressions to detectable linguistic patterns that strip them of their meaning. Addressing it will require new approaches to what counts as detectable knowledge in AI models, new forms of collaboration between students, teachers and developers, and new frameworks for evaluating whether AI tools are supporting the full range of learning goals that justice-centered science education demands.

2.5 Design Shapes What Teachers Notice and Act On

A recurring theme across the three chapters is that the design of the AI tool itself shapes what teachers are able to notice in student thinking (Luna, 2018; Wilkerson et al., 2016) and what they can act on within the real constraints of teaching (Voogt et al., 2015). In the early iterations of this work, teachers accessed student dialog data primarily through the progress review (e.g., how many percentage of students finished the dialog), reading through selected student dialogs in RPP workshops after they finished the unit, discussing patterns qualitatively, and making revision decisions collaboratively. This approach produced valuable insights but required significant time and researcher mediation. What teachers found most difficult was quickly identifying whole-class patterns: which ideas were most frequently expressed, which were missing, and where the class as a whole needed more instructional support.

Ongoing work in the RPP is addressing this through the development of a teacher-facing dashboard (Wiley, et. al., 2024) that surfaces aggregate patterns such as mean KI scores and idea frequency counts organized by idea type alongside individual student responses. Early classroom observations of teachers using the dashboard documented teachers noticing whole-class patterns and responding within the same instructional period: one teacher acknowledged students' dialog contributions publicly, building motivation; another identified a frequently expressed idea from the frequency data and built on it the following class. The design lesson is that tools designed to support teacher noticing must attend not only to what information is available but to what form makes whole-class patterns visible and usable given the pace and complexity of real classrooms.

3 Limitations

Two limitations constrain the conclusions of this dissertation. First, both focal teachers across Chapters 3 and 4 are long-standing RPP members with deep familiarity with KI frameworks, co-design processes, and their students' communities. Their implementations represent what is possible under favorable conditions. Whether the RPP iterative co-design structures and activities produce comparable results with teachers who have less RPP experience, less theoretical grounding, or less institutional support for customization remains an important open question.

Second, the annotation process that makes the NLP models work is labor-intensive and difficult to scale. Each new idea added to the rubric requires re-annotation of existing datasets, retraining of the model, and renewed inter-rater reliability work between researchers. This dissertation required multiple full annotation cycles involving hundreds of hours of researcher labor. As the dialogs are extended to new topics, new communities, and new languages, this bottleneck will become increasingly limiting. Approaches that reduce annotation burden, including few-shot learning, zero-shot classification, and teacher-facing annotation tools that distribute the labor more broadly, deserve serious attention in future work.

4 Implications

The most important implication of this dissertation for AI and NLP developers is to include teachers as co-designers throughout the full arc of tool development, not just at validation and implementation. What teachers contribute is valuable knowledge that developers cannot access any other way: language patterns their specific students use to express ideas, the cultural and community contexts that give those ideas meaning, and the pedagogical priorities that determine which ideas are worth detecting and responding to in the first place. In the food justice work, teachers identified ideas tied to students' lived experiences of food access in their communities that were not visible in the training data until teachers named them. No amount of technical sophistication in model architecture can substitute for that kind of local, relational knowledge.

A second implication concerns what counts as model effectiveness. The field has tended to evaluate NLP models in education primarily on accuracy metrics on a fixed rubric such as F1 scores, precision, and recall. This dissertation suggests that rubric quality is at least as important as model accuracy, and that rubric quality can only be evaluated with reference to the pedagogical goals of the teachers who will use the tool. Evaluating AI tools for education requires building teacher judgment into the evaluation process itself, not just into the design process.

For science teachers, the most important implication of this dissertation is that their expertise has the power to shape what AI can do in their classrooms, but only if they engage with that expertise intentionally. Amy's case across Chapters 3 and 4 illustrates what this looks like in practice: she did not treat the AI dialog as a finished tool to be implemented. She treated it as a starting point that her knowledge of her students could improve. She revised guidance language, designed complementary activities, and used dialog data to inform her instruction, and each of those contributions made the dialog more effective for her students and, through the RPP feedback loop, for students in other classrooms as well.

One concrete example of what teacher agency in AI-enhanced classrooms can look like comes from ongoing work with Amy after the studies documented in this dissertation. Amy designed a classroom annotation activity in which students, after completing the pretest dialog, collaboratively annotated anonymized peer responses from the AI dialog by underlining important ideas, boxing key vocabulary, writing margin notes for improvement, and marking connections to their own ideas. Students engaged deeply with the activity, developed understanding of how the dialog worked, and produced significantly longer and more integrated responses at posttest. This example suggests that teachers who understand how AI dialogs function, including what they detect, how they respond, and what they cannot see, are better positioned to design instruction that works with the dialog rather than around it. Building that understanding should be a goal of professional development for teachers working in AI-enhanced science classrooms.

5 Conclusion

I remember a moment in an RPP workshop when a teacher looked at a student's dialog response and said: “this is a good entry point. I would ask more follow-up questions here. I would build on this.” That moment of teacher noticing is specific, pedagogically grounded, rooted in years of knowing students like the one who wrote that response. To make that noticing count in an AI dialog, I needed to define the ideas precisely enough to train a model, reach inter-rater reliability with other researchers on thousands of annotated examples, wait for computer scientists to train the models, design and revise guidance for each detectable idea, and then test whether the new guidance actually helped students think further in real classrooms. That process, from a teacher's insight to a deployed dialog, is what this dissertation documents. It is slow and collaborative by design, because the knowledge it builds into the tool cannot be acquired any other way.

The classrooms documented in this dissertation show what this work makes possible. Students began with intuitive ideas about growing plants, and the photosynthesis AI dialog honored those ideas as genuine starting points rather than errors to correct. Through the photosynthesis and ecosystem lessons, students traced how energy moves from sunlight through plants into the organisms and communities that depend on them, grounded in a school garden they could see, design, and tend. That same garden then became the entry point for reasoning about why some communities can grow or access fresh food and others cannot, and for designing action to change it. The food justice AI dialog detected not just science mechanisms but community observations, structural reasoning, and calls for action, which are the ideas shaped by students' own experiences of the neighborhoods, grocery stores, and school gardens they knew. Throughout, the food justice perspective was never separate: students asking why some plants thrive and others do not were already asking why some communities can grow and access fresh food while others cannot. KI and JCSP were combined on the same inquiry, making science accessible and justice-centered at every step.

Designing AI dialogs that genuinely honor the full range of student thinking, including the community-based, culturally grounded, and justice-oriented ideas that standardized tools tend to miss, requires sustained partnerships, iterative refinement, and a willingness to let teacher expertise reshape what the technology does rather than simply validate what it already does. The classrooms documented in this dissertation, where students reasoned about ecosystems in their school gardens and grocery stores they knew from their own communities, and where teachers designed instruction that treated student knowledge as a resource rather than a deficit, show what becomes possible when that work is done carefully and collaboratively. This dissertation offers the field actionable design guidelines and empirical evidence for building AI dialogs that are worthy of the students and teachers they are designed to serve.

References

Akgun, S., & Greenhow, C. (2022). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *Ai and Ethics*, *2*(3), 431–440. https://doi.org/10.1007/s43681-021-00096-7

Aleven, V. A., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science*, *26*, 147–179.

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. https://doi.org/10.3102/0162373716659929

Bang, M., & Medin, D. (2010). Cultural processes in science education: Supporting the navigation of multiple epistemologies. *Science Education*, *94*(6), 1008–1026. https://doi.org/10.1002/sce.20392

Barton, A. C., & Tan, E. (2009). Funds of knowledge and discourses and hybrid space. *Journal of Research in Science Teaching*, *46*(1), 50–73. https://doi.org/10.1002/tea.20269

Beltagy, I., Lo, K., & Cohan, A. (2019). *SciBERT: A Pretrained Language Model for Scientific Text* (arXiv:1903.10676). arXiv. https://doi.org/10.48550/arXiv.1903.10676

Bichler, S., Gerard, L., Bradford, A., & Linn, M. C. (2021). Designing a remote professional development course to support teacher customization in science. *Computers in Human Behavior*, *123, 106814*. https://doi.org/10.1016/j.chb.2021.106814

Billings, K., & Linn, M. C. (2024). A Case Study of Activity Structures and Resources that Support Teacher Agency and Leadership in an RPP. In R. Lindgren, T. I. Asino, E. A. Kyza, C. K. Looi, D. T. Keifert, & E. Suárez (Eds.), *Proceedings of the 18th International Conference of the Learning Sciences—ICLS 2024* (pp. 682–689). International Society of the Learning Sciences. https://repository.isls.org//handle/1/11156

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of “Bias” in NLP. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Bradford, A., Bichler, S., & Linn, M. C. (2021). Designing a Workshop to Support Teacher Customization of Curricula. In E. de Vries, J. Ahn, & Y. Hod (Eds.), *Proceedings of the 15th International Conference of the Learning Sciences – ICLS 2021* (pp. 100–115). International Society of the Learning Sciences.

Bradford, A., Gerard, L., Tate, E., Li, R., & Linn, M. C. (2023). Incorporating investigations of environmental racism into middle school science. *Science Education*, *107*(6), 1628-1654.

Bradford, A., Li, W., Steimel, K., Riordan, B., & Linn, M. C. (2023). Adaptive dialog to support student understanding of climate change mechanism and who is most impacted. In P. Blikstein, J. Van Aalst, R. Kizito, & K. Brennan (Eds.), *Proceedings of the 17th International Conference of the Learning Sciences—ICLS 2023* (pp. 816–823). International Society of the Learning Sciences. https://doi.org/10.22318/icls2023.681776

Bradford, A., Steimel, K., Riordan, B., & Linn, M. (2024, June). Building Robust Content Scoring Models for Student Explanations of Social Justice Science Issues. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 450-458).

Chowdhary, K. (2020). Natural language processing. In *Fundamentals of artificial intelligence* (pp. 603–649). Springer.

Clark, D. B., & Sampson, V. D. (2007). Personally-Seeded Discussions to Scaffold Online Argumentation. *International Journal of Science Education*, *29*(3), 253–277. https://doi.org/10.1080/09500690600560944

Coburn, C. E., & Penuel, W. R. (2016). Research–Practice Partnerships in Education: Outcomes, Dynamics, and Open Questions. *Educational Researcher*, *45*(1), 48–54. https://doi.org/10.3102/0013189X16631750

Condor, A., Litster, M., & Pardos, Z. (2021). Automatic Short Answer Grading with SBERT on Out-of-Sample Questions. *International Educational Data Mining Society*.

Condor, A., Pardos, Z., & Linn, M. (2022, July). Representing scoring rubrics as graphs for automatic short answer grading. In *International Conference on Artificial Intelligence in Education* (pp. 354-365). Cham: Springer International Publishing.

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

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

diSessa, A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, *10*(2–3), 105–225.

diSessa, A. A. (1988). Knowledge in pieces. In G. Forman & P. Pufall (Eds.), *Constructivism in the computer age* (pp. 49–70). Lawrence Erlbaum Associates.

diSessa, A. A. (2018). A Friendly Introduction to "Knowledge in Pieces": Modeling Types of Knowledge and Their Roles in Learning. In G. Kaiser, H. Forgasz, M. Graven, A. Kuzniak, E. Simmt, & B. Xu (Eds.), *Invited Lectures from the 13th International Congress on Mathematical Education* (pp. 65–84). Springer International Publishing. https://doi.org/10.1007/978-3-319-72170-5_5

diSessa, A. A., & Sherin, B. L. (1998). What changes in conceptual change? *International Journal Of Science Education*, *20*(10), 1155–1191.

Donnelly, D. F., Namdar, B., Vitale, J. M., Lai, K., & Linn, M. C. (2016). Enhancing student explanations of evolution: Comparing elaborating and competing theory prompts. *Journal of Research in Science Teaching*, *53*(9), 1341–1363. https://doi.org/10.1002/tea.21331

Durall Gazulla, E., Martins, L., & Fernández-Ferrer, M. (2023). Designing learning technology collaboratively: Analysis of a chatbot co-design. *Education and Information Technologies*, *28*(1), 109–134. https://doi.org/10.1007/s10639-022-11162-w

Fischer, F., Kollar, I., Stegmann, K., & Wecker, C. (2013). Toward a script theory of guidance in computer-supported collaborative learning. *Educational psychologist*, 48(1), 56-66.

Freire, P. (1970). The adult literacy process as cultural action for freedom. *Harvard educational review*, 40(2), 205-225.

Gerard, L., Bichler, S., Bradford, A., Linn, M. C., Steimel, K., & Riordan, B. (2022). Designing an Adaptive Dialogue to Promote Science Understanding. *Proceedings of the 16th International Conference of the Learning Sciences - ICLS 2022*. https://par.nsf.gov/biblio/10330141-designing-adaptive-dialogue-promote-science-understanding

Gerard, L., Bradford, A., Debarger, A., Wiley, K., & Linn, M. C. (2022). Cultivating teacher efficacy for social justice in science. *Science Scope*, 45(5), 35-43.

Gerard, L., Bradford, A., & Linn, M. C. (2022). Supporting Teachers to Customize Curriculum for Self-Directed Learning. *Journal of Science Education and Technology*, *31*(5), 660–679. https://doi.org/10.1007/s10956-022-09985-w

Gerard, L. F., Spitulnik, M., & Linn, M. C. (2010). Teacher use of evidence to customize inquiry science instruction. *Journal of Research in Science Teaching*, *47*(9), 1037–1063. https://doi.org/10.1002/tea.20367

Gerard, L. F., Varma, K., Corliss, S. B., & Linn, M. C. (2011). Professional Development for Technology-Enhanced Inquiry Science. *Review of Educational Research*, *81*(3), 408–448. https://doi.org/10.3102/0034654311415121

Gerard, L., Holtman, M., Riordan, B., & Linn, M. C. (2025). Impact of an adaptive dialog that uses natural language processing to detect students' ideas and guide knowledge integration. *Journal of Educational Psychology*, *117*(1), 63–87. https://doi.org/10.1037/edu0000902

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. https://doi.org/10.1007/s10972-016-9455-6

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

Gerard, L., Linn, M. C., & Holtmann, M. (2024). A comparison of responsive and general guidance to promote learning in an online science dialog. *Education Sciences*, *14*(12), 1383.

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

Gerard, L. F., Ryoo, K., McElhaney, K. W., Liu, O. L., Rafferty, A. N., & Linn, M. C. (2016). Automated guidance for student inquiry. *Journal of Educational Psychology*, *108*(1), 60–81. https://doi.org/10.1037/edu0000052

Graesser, A. C. (2016). Conversations with AutoTutor Help Students Learn. *International Journal of Artificial Intelligence in Education*, *26*(1), 124–132. https://doi.org/10.1007/s40593-015-0086-4

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 180–192. https://doi.org/10.3758/BF03195563

He, P., Gao, J., & Chen, W. (2021). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543.*

Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4), 470-497.

Holtmann, M., Gerard, L., Li, W., Linn, M. C., Riordan, B., & Steimel, K. (2023). How Does an Adaptive Dialog Based on Natural Language Processing Impact Students from Distinct Language Backgrounds? In P. Blikstein, J. Van Aalst, R. Kizito, & K. Brennan (Eds.), *Proceedings of the 17th International Conference of the Learning Sciences—ICLS 2023* (pp. 1350–1353). International Society of the Learning Sciences. https://doi.org/10.22318/icls2023.921177

Hutchins, N. M., & Biswas, G. (2024). Co-designing teacher support technology for problem-based learning in middle school science. *British Journal of Educational Technology*, *55*(3), 802–822. https://doi.org/10.1111/bjet.13363

Ibourk, A., Wagner, L., Morrison, D., Young, S., & Milledge, J. (2024). Community gardens as places for ecological caring in action. *Science and Children*, 61(2), 83–87. https://doi.org/10.1080/00368148.2024.2315673

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking: From childhood to adolescence.* (A. Parsons & S. Milgram, Trans.). Basic Books. https://doi.org/10.1037/10034-000

Jacobs, V. R., Lamb, L. L., & Philipp, R. A. (2010). Professional noticing of children's mathematical thinking. *Journal for research in mathematics education*, 41(2), 169-202.

Jayathirtha, G., Chapman, G., & Goode, J. (2025). Questioning the why and the how: Collective transformative agency of experienced teachers co-designing a justice-oriented high school introductory computing program. *Journal of Research on Technology in Education*, *57*(1), 84–109. https://doi.org/10.1080/15391523.2024.2399278

Jurenka, I., Kunesch, M., McKee, K. R., Gillick, D., Zhu, S., Wiltberger, S., Phal, S. M., Hermann, K., Kasenberg, D., Bhoopchand, A., Anand, A., Pîslar, M., Chan, S., Wang, L., She, J., Mahmoudieh, P., Rysbek, A., Ko, W.-J., Huber, A., … Ibrahim, L. (2024). *Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach* (arXiv:2407.12687). arXiv. http://arxiv.org/abs/2407.12687

Kali, Y. (2006). Collaborative knowledge building using the Design Principles Database. *International Journal of Computer-supported collaborative learning*, 1(2), 187-201.

Kang, H., Windschitl, M., Stroupe, D., & Thompson, J. (2016). Designing, launching, and implementing high quality learning opportunities for students that advance scientific thinking. *Journal of Research in Science Teaching*, *53*(9), 1316–1340. https://doi.org/10.1002/tea.21329

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., & Gurevych, I. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In D. Zhao (Ed.), *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* (pp. 5–9). Association for Computational Linguistics. https://aclanthology.org/C18-2002/

Kubsch, M., Krist, C., & Rosenberg, J. M. (2023). Distributing epistemic functions and tasks—A framework for augmenting human analytic power with machine learning in science education research. *Journal of Research in Science Teaching*, *60*(2), 423–447. https://doi.org/10.1002/tea.21803

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Review of Educational Research*, *86*(1), 42–78. https://doi.org/10.3102/0034654315581420

Kuiper, J. (1994). Student ideas of science concepts: alternative frameworks?. *International Journal of Science Education*, 16(3), 279-292.

Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American educational research journal*, 32(3), 465-491.

Lee, H.-S., Gweon, G.-H., Lord, T., Paessel, N., Pallant, A., & Pryputniewicz, S. (2021). Machine Learning-Enabled Automated Feedback: Supporting Students' Revision of Scientific Arguments Based on Data Drawn from Simulation. *Journal of Science Education and Technology*, *30*(2), 168–192. https://doi.org/10.1007/s10956-020-09889-7

Lee, O., & Buxton, C. (2011). Engaging culturally and linguistically diverse students in learning science. *Theory Into Practice*, *50*(4), 277–284.

Li, W. (2025). Applying natural language processing adaptive dialogs to promote knowledge integration during instruction. *Education Sciences*, 15(2), 207. https://doi.org/10.3390/educsci15020207

Li, W., Chang, H.-Y., Bradford, A., Gerard, L., & Linn, M. C. (2024). Combining Natural Language Processing with Epistemic Network Analysis to Investigate Student Knowledge Integration within an AI Dialog. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-024-10176-y

Li, W., Liao, Y., Steimel, K., Bradford, A., Gerard, L., & Linn, M. (2024). Teacher-informed Expansion of an Idea Detection Model for a Knowledge Integration Assessment. *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, 447–450. https://doi.org/10.1145/3657604.3664687

Li, W., Lim-Breitbart, J., Bradford, A., Linn, M. C., Riordan, B., & Steimel, K. (2023). Explaining Thermodynamics: Impact of an Adaptive Dialog based on a Natural Language Processing Idea Detection Model. In P. Blikstein, J. Van Aalst, R. Kizito, & K. Brennan (Eds.), *Proceedings of the 17th International Conference of the Learning Sciences—ICLS 2023* (pp. 1306–1309). International Society of the Learning Sciences. https://doi.org/10.22318/icls2023.199424

Lincoln, Y. S., & Guba, E. (1985). *Naturalistic Inquiry*. SAGE Publications.

Linn, M. C. (2006). The Knowledge Integration Perspective on Learning and Instruction. In R. K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (1 edition, pp. 243–264). Cambridge University Press.

Linn, M. C., Clark, D., & Slotta, J. D. (2003). WISE design for knowledge integration. *Science Education*, *87*(4), 517–538. https://doi.org/10.1002/sce.10086

Linn, M. C., Davis, E. A., & Bell, P. (2004). Specific Design Principles: Elaborating the Scaffolded Knowledge Integration Framework. In M. C. Linn, E. A. Davis, & P. Bell (Eds.), *Internet Environments for Science Education* (pp. 315–340). Lawrence Erlbaum Associates.

Linn, M. C., Donnelly-Hermosillo, D., & Gerard, L. F. (2023). Synergies between learning technologies and learning sciences: Promoting equitable secondary science education. In N. Lederman, D. Zeidler, & J. Lederman (Eds.), *Handbook of Research on Science Education: Vol. III* (1st ed., pp. 447–498). Routledge Press.

Linn, M. C., & Eylon, B.-S. (2011). *Science Learning and Instruction: Taking Advantage of Technology to Promote Knowledge Integration*. Routledge. https://doi.org/10.4324/9780203806524

Linn, M. C., & Gerard, L. (2024). Assessing and Guiding Student Science Learning with Pedagogically Informed Natural Language Processing. In X. Zhai & J. Krajcik (Eds.), *Uses of Artificial Intelligence in STEM Education* (pp. 59–88). Oxford University Press. https://doi.org/10.1093/oso/9780198882077.003.0004

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

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

Lowell, B. R., Fogelman, S. E., & McNeill, K. L. (2024). Organizational sensemaking during curriculum implementation: The dilemma of agency, role of collaboration, and importance of discipline-specific leadership. *Science Education*, *108*(5), 1448–1473. https://doi.org/10.1002/sce.21885

Luna, M. J. (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. https://doi.org/10.1080/07370008.2018.1496919

Mangal, M., & Pardos, Z. A. (2024). Implementing equitable and intersectionality-aware ML in education: A practical guide. *British Journal of Educational Technology*, 55(5), 2003-2038.

Matuk, C., Gerard, L., Lim-Breitbart, J., & Linn, M. (2016). Gathering requirements for teacher tools: Strategies for empowering teachers through co-design. *Journal of Science Teacher Education*, 27(1), 79-110.

Matuk, C., McElhaney, K. W., Chen, J. K., Lim-Breitbart, J., Kirkpatrick, D., & Linn, M. C. (2016). Iteratively Refining a Science Explanation Tool Through Classroom Implementation and Stakeholder Partnerships. *International Journal of Designs for Learning*, *7*(2). https://doi.org/10.14434/ijdl.v7i2.20203

Matuk, C., & Linn, M. C. (2018). Why and how do middle school students exchange ideas during science inquiry?. *International Journal of Computer-Supported Collaborative Learning*, 13(3), 263-299.

Matuk, C., & Linn, M. C. (2023). Students' perceptions of the impacts of peer ideas in inquiry learning. *Instructional Science*, 51(1), 65-102.

McKenney, S., & Reeves, T. (2018). *Conducting educational design research*. Routledge.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). Qualitative data analysis: A methods sourcebook (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Morales-Doyle, D. (2017). Justice-centered science pedagogy: A catalyst for academic achievement and social transformation. *Science Education*, *101*(6), 1034–1060. https://doi.org/10.1002/sce.21305

Morales-Doyle, D. (2018). Students as curriculum critics: Standpoints with respect to relevance, goals, and science. *Journal of Research in Science Teaching*, *55*(5), 749-773.

Myers, M. C., & Wilson, J. (2023). Evaluating the Construct Validity of an Automated Writing Evaluation System with a Randomization Algorithm. *International Journal of Artificial Intelligence in Education*, *33*(3), 609–634. https://doi.org/10.1007/s40593-022-00301-6

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. National Academies Press. http://www.nextgenscience.org/next-generation-science-standards

Nyaaba, M., & Zhai, X. (2024). *Developing Custom GPTs for Education: Bridging Cultural and Contextual Divide in Generative AI* (SSRN Scholarly Paper 5074403). Social Science Research Network. https://doi.org/10.2139/ssrn.5074403

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and Family: A Review of 17 Years of Natural Language Tutoring. *International Journal of Artificial Intelligence in Education*, *24*(4), 427–469. https://doi.org/10.1007/s40593-014-0029-5

Otero, V. K. (2006). Moving Beyond the "Get it or Don't" Conception of Formative Assessment. *Journal of Teacher Education*, *57*(3), 247–255. https://doi.org/10.1177/0022487105285963

Paladines, J., & Ramirez, J. (2020). A Systematic Literature Review of Intelligent Tutoring Systems With Dialogue in Natural Language. *IEEE Access*, *8*, 164246–164267. IEEE Access. https://doi.org/10.1109/ACCESS.2020.3021383

Pardos, Z. A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *Plos one*, 19(5), e0304013.

Pardos, Z. A., Tang, M., Anastasopoulos, I., Sheel, S. K., & Zhang, E. (2023). OATutor: An open-source adaptive tutoring system and curated content library for learning sciences research. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–17).

Paris, D., & Alim, H. S. (2017). Culturally Sustaining Pedagogies: Teaching and Learning for Justice in a Changing World. In *Teachers College Press*. Teachers College Press.

Penuel, W., Gallagher, L., & Moorthy, S. (2011). Preparing Teachers to Design Sequences of Instruction in Earth Systems Science: A Comparison of Three Professional Development Programs. *American Educational Research Journal - AMER EDUC RES J*, *48*, 996–1025. https://doi.org/10.3102/0002831211410864

Penuel, W., Krumm, A., Pazera, C., Singleton, C., Allen, A., & Deverel-Rico, C. (2023). Belonging in science classrooms: Investigating its relation to students' contributions and influence in knowledge building. *Journal of Research in Science Teaching*, *61*, 228–252. https://doi.org/10.1002/tea.21884

Penuel, W., Roschelle, J., & Shechtman, N. (2007). DESIGNING FORMATIVE ASSESSMENT SOFTWARE WITH TEACHERS: AN ANALYSIS OF THE CO-DESIGN PROCESS. *Research and Practice in Technology Enhanced Learning*, *2*, 51–74. https://doi.org/10.58459/rptel.2007.251-74

Philip, T. M., & Azevedo, F. S. (2017). Everyday science learning and equity: Mapping the contested terrain. *Science Education*, 101(4), 526-532.

Philip, T. M., Pham, J. H., Scott, M., & Cortez, A. (2022). Intentionally Addressing Nested Systems of Power in Schooling through Teacher Solidarity Co-Design. *Cognition and Instruction*, *40*(1), 55–76. https://doi.org/10.1080/07370008.2021.2010208

Puntambekar, S., Dey, I., Gnesdilow, D., Passonneau, R. J., & Kim, C. (2023). *Examining the Effect of Automated Assessments and Feedback on Students' Written Science Explanations*. https://repository.isls.org//handle/1/10060

Remillard, J. T. (2000). Can curriculum materials support teachers' learning? Two fourth-grade teachers' use of a new mathematics text. *The elementary school journal*, 100(4), 331-350.

Reza, M., Anastasopoulos, I., Bhandari, S., & Pardos, Z. A. (2025). PromptHive: Bringing subject matter experts back to the forefront with collaborative prompt engineering for educational content creation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–22).

Riordan, B., Cahill, A., Chen, J. K., Wiley, K., Bradford, A., Gerard, L., & & Linn, M. C. (2020). Identifying NGSS-Aligned Ideas in Student Science Explanations. *Workshop on Artificial Intelligence for Education (AI4EDU@AAAI)*. https://par.nsf.gov/biblio/10184621-identifying-ngss-aligned-ideas-student-science-explanations

Riordan, B., Bichler, S., Bradford, A., Chen, J. K., Wiley, K., Gerard, L., & Linn, M. C. (2020, July). 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* (pp. 135-144).

Riordan, B., Cahill, A., Chen, J. K., Wiley, K., Bradford, A., Gerard, L., & & Linn, M. C. (2020). Identifying NGSS-Aligned Ideas in Student Science Explanations. *Workshop on Artificial Intelligence for Education (AI4EDU@AAAI)*. https://par.nsf.gov/biblio/10184621-identifying-ngss-aligned-ideas-student-science-explanations

Riordan, B., Wiley, K., Chen, J. K., Bradford, A., Bichler, S., Mulholland, M., & Gerard, L. F. (2020, April). Automated scoring of science explanations for multiple NGSS dimensions and knowledge integration. In *Annual Meeting of the American Educational Research Association (AERA)*.

Ritchhart, R., & Perkins, D. (2008). Making thinking visible. *Educational leadership*, 65(5), 57.
Rivera Maulucci, M. S., Brown, B. A., Grey, S. T., & Sullivan, S. (2014). Urban middle school students' reflections on authentic science inquiry. *Journal of Research in Science Teaching*, *51*(9), 1119–1149. https://doi.org/10.1002/tea.21167
Rodriguez, G. M. (2013). Power and Agency in Education: Exploring the Pedagogical Dimensions of Funds of Knowledge. *Review of Research in Education*, *37*, 87–120.
Rosebery, A. S., Warren, B., & Tucker-Raymond, E. (2016). Developing interpretive power in science teaching. *Journal of Research in Science Teaching*, *53*(10), 1571–1600. https://doi.org/10.1002/tea.21267
Roschelle, J., & Penuel, W. (2006). *Co-design of innovations with teachers: Definition and dynamics*. *2*, 606–612.
Ryoo, K., & Linn, M. C. (2012). Can dynamic visualizations improve middle school students' understanding of energy in photosynthesis?. *Journal of Research in Science Teaching*, 49(2), 218-243.
Ryoo, K., & Linn, M. C. (2015). Designing and Validating Assessments of Complex Thinking in Science. *Theory Into Practice*, *54*(3), 238–254. https://doi.org/10.1080/00405841.2015.1044374
Sandoval, W. A., & Bell, P. (2004). Design-Based Research Methods for Studying Learning in Context: Introduction. *Educational Psychologist*, *39*(4), 199–201. https://doi.org/10.1207/s15326985ep3904_1
Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The Risk of Racial Bias in Hate Speech Detection. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1668–1678). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1163
Sayary, A. M. A. E., Forawi, S. A., & Mansour, N. (2015). STEM education and problem-based learning. In *The Routledge International Handbook of Research on Teaching Thinking*. Routledge.
Schulz, C., Eger, S., Daxenberger, J., Kahse, T., & Gurevych, I. (2018). *Multi-Task Learning for Argumentation Mining in Low-Resource Settings* (arXiv:1804.04083). arXiv. https://doi.org/10.48550/arXiv.1804.04083
Schulz, C., Meyer, C. M., & Gurevych, I. (2019). Challenges in the Automatic Analysis of Students' Diagnostic Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), Article 01. https://doi.org/10.1609/aaai.v33i01.33016974
Severance, S., Penuel, W. R., Sumner, T., & Leary, H. (2016). Organizing for Teacher Agency in Curricular Co-Design. *Journal of the Learning Sciences*, *25*(4), 531–564. https://doi.org/10.1080/10508406.2016.1207541
Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, *78*(1), 153–189.
Smith, J. P., DiSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences*, 3(2), 115–163.
Songer, N. B., & Linn, M. C. (1991). How do students' views of science influence knowledge integration?. *Journal of Research in Science Teaching*, 28(9), 761-784.
Strauss, A, & Corbin, J. (1998). *Basics of qualitative research: techniques and procedures for developing grounded theory* (2nd ed.). Sage.
Svihla, V., Reeve, R., Sagy, O., & Kali, Y. (2015). A fingerprint pattern of supports for teachers' designing of technology-enhanced learning. *Instructional Science*, *43*(2), 283–307. https://doi.org/10.1007/s11251-014-9342-5
Tansomboon, C., Gerard, L. F., Vitale, J. M., & Linn, M. C. (2017). Designing Automated Guidance to Promote Productive Revision of Science Explanations. *International Journal of Artificial Intelligence in Education*, *27*(4), 729–757. https://doi.org/10.1007/s40593-017-0145-0
Valencia, R. R. (1997). *The evolution of deficit thinking: Educational thought and practice* (pp. xvii, 270). The Falmer Press/Taylor & Francis.

Vitale, J. M., McBride, E., & Linn, M. C. (2016). Distinguishing complex ideas about climate change: Knowledge integration vs. specific guidance. *International Journal of Science Education*, *38*(9), 1548–1569. https://doi.org/10.1080/09500693.2016.1198969
Voogt, J., Laferriere, T., Breuleux, A., Itow, R., Hickey, D., & Mckenney, S. (2015). Collaborative design as a form of professional development. *Instructional Science*, *43*. https://doi.org/10.1007/s11251-014-9340-7
Vossoughi, S., & Bevan, B. (2014). Making and Tinkering: A Review of the Literature. *National Research Council Committee on Out of School Time STEM*, 1–55.
Vygotsky, L. S., & Cole, M. (1978). *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press.
Walker, E., Rummel, N., & Koedinger, K. R. (2011). Using Automated Dialog Analysis to Assess Peer Tutoring and Trigger Effective Support. In S. B. G. Biswas & A. M. J. Kay (Eds.), *Proceedings of the 10th International Conference on Artificial Intelligence in Education* (pp. 385–393). Springer.
Walkoe, J., Wilkerson, M., & Elby, A. (2017). Technology-mediated teacher noticing: A goal for classroom practice, tool design, and professional development. Philadelphia, PA: International Society of the Learning Sciences.
Wiley, K., Bradford, A., & Linn, M. C. (2019). Supporting Collaborative Curriculum Customizations Using the Knowledge Integration Framework. *Computer-Supported Collaborative Learning*, *1*, 480–487. https://par.nsf.gov/biblio/10106811-supporting-collaborative-curriculum-customizations-using-knowledge-integration-framework
Wiley, K., Dimitriadis, Y., & Linn, M. (2024). A human-centred learning analytics approach for developing contextually scalable K-12 teacher dashboards. *British Journal of Educational Technology*, *55*(3), 845–885. https://doi.org/10.1111/bjet.13383
Wilkerson, M. H., Andrews, C., Shaban, Y., Laina, V., & Gravel, B. E. (2016). What's the technology for? Teacher attention and pedagogical goals in a modeling-focused professional development workshop. *Journal of Science Teacher Education*, 27(1), 11-33.
Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are We There Yet? - A Systematic Literature Review on Chatbots in Education. *Frontiers in Artificial Intelligence*, *4*, 654924. https://doi.org/10.3389/frai.2021.654924
Xia, Q., Li, W., Yang, Y., Weng, X., & Chiu, T. K. F. (2025). A systematic review and meta-analysis of the effectiveness of Generative Artificial Intelligence (GenAI) on students' motivation and engagement. *Computers and Education: Artificial Intelligence*, *9*, 100455. https://doi.org/10.1016/j.caeai.2025.100455
Yin, R. K. (2018). *Case study research and applications* (Vol. 6). Thousand Oaks, CA: Sage.
Zhai, X., Haudek, K. C., Stuhlsatz, M. A., & Wilson, C. (2020). Evaluation of construct-irrelevant variance yielded by machine and human scoring of a science teacher PCK constructed response assessment. *Studies in Educational Evaluation*, 67, 100916.
Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education*, *56*(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757
Zhu, M., Liu, O. L., & Lee, H.-S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, *143*, 103668. https://doi.org/10.1016/j.compedu.2019.103668

Appendix

Table 1 KI Scoring Rubric, Idea Rubric and adaptive guidance for the AI Dialog at Iteration 1 (Original idea rubric).

Dialog prompt: Please explain your ideas about how animals get and use energy from the sun to survive. Be sure to explain how energy and matter move and how energy and matter change

KI	Description	Idea and description	Adaptive Guidance
1	Irrelevant / off-topic	Off topic ideas (e.g., I don't know.)	Can you tell me more about this idea or another one in your explanation? I am still learning about student ideas to become a better thought partner.
2	**No link** Incomplete / vague / inaccurate ideas	4-EngCreate: Energy is created not transformed/transferred(e.g., chemical energy is created by plants)	Cannot be accurately detected
		5-Eng2Mat: Plants transform (convert/change/turn) light energy into glucose/sugar/food OR turn/transform glucose into energy	Cannot be accurately detected
		11-AnimDirUse: Animals directly use the Sun's energy for vitamins and keeping warm.	Interesting idea! Can animals live without sunlight? How does the animal use energy from the sun?
		12-AnimFood: Animals eat plants [focused on animals eat plants as food and NOT about energy such as sun grows plants which are energy source for animals]	Interesting idea! How do plants and animals use energy from the sun differently?
3	**Partial link** Accurate idea(s), but isolated (conclusion only, explanation/ evidence only)	1-PhotoRec: CO_2, H_2O or both as reactants of photosynthesis	Nice thinking. You mentioned the inputs of photosynthesis. What are the outputs of this process?
		2-PhotoProd: glucose [or sugar or food] or oxygen as a product	Interesting idea about sugar as a product in this process. How are the products of photosynthesis useful for animals?
		3-Photo: plant uses energy from the sun to do photosynthesis	Nice thinking of photosynthesis. How are the products of photosynthesis useful for animals?
		3a-PhotoChem: Energy from the sun transforms into another type of energy [kinetic/chemical/usable] during photosynthesis	Interesting idea about how plants transform light energy to usable energy. How does the energy get to animals?
		6a-PltStore: Plants store energy in glucose.	Cannot be accurately detected
		6-PltCellResp: Plant releases energy from glucose/food for: growth, energy, repair, seed production	Cannot be accurately detected
		8-EngTrans: Energy from the sun gets to animals when they eat plants	Nice thinking! You talked about energy transfer. Can you tell me more about how animals use the energy?
		9-AnimCellResp: Animal uses cellular respiration to release energy	Nice thinking! Can you tell me more about how the animal release the energy they get from the plant?
		10-AnimGrw: Animal uses glucose/food for energy, repair, growth, to move	Cannot be accurately detected
4	**Single link**: One scientifically complete and valid connection between KI 3 ideas		

5	**Multiple links**: Two or more scientifically complete and valid connections between KI 3 ideas

Appendix Table 2. Revised Idea Rubric and adaptive guidance for the Energy Story AI Dialog at Iteration 3 (Green colored ideas are normative ideas; blue colored ideas are vague/broad; orange colored ideas are intuitive)

#	Idea description	Student Example	Guidance
	Nonscorable	I don't know	How does the sun help animals live? How does the energy from the sun help animals live?
1	Reactants of photosynthesis (e.g., water, carbon dioxide)	"plants need water, CO2, and the sun to do photosynthesis."	Nice thinking! What happens after photosynthesis? Good progress! How does photosynthesis help the animal get energy from plants?
2	Products of photosynthesis (e.g., glucose, oxygen)	"the sun makes the plant do photosynthesis which produces oxygen" "The plant use photosynthesis to made Sugar and oxygen"	Interesting ideas! How do animals use the glucose after they get it? Great! How are the products useful for animals?
3	Energy from the sun helps plants grow	"Sun helps plants to grow and many animals eat plants" "the energy of the sun helps the plants grow"	Interesting ideas! How does the plant use the solar energy from the sun? Interesting! How does the plant take in energy from the sun to grow?
3a	Energy from sun transforms into another type of energy [kinetic /chemical / usable *NOT thermal]	"plants use chlorophyll to photosynthesize the Sun's energy into plant energy"	Nice thinking! How is the energy from the sun changed into usable energy for plants? Great! What are the processes that change solar energy into usable energy for animals?
3b	Plants get energy from the sun through photosynthesis (keyword or describe chem reaction)	"The animals use sun energy from the sun to survive by using the process of photosynthesis and eating plants who also use the energy from the sun."	Nice thinking! How does the energy from photosynthesis make its way to animals? Good progress! How do the products of photosynthesis help animals live?
4a	Energy is created by plants	"Plants create energy, which is then consumed by animals."	Thanks for sharing with me! Where does the energy originally come from? Interesting! How do plants get energy from the sun, differently than animals?
4b	Plants turn light energy into matter (glucose/sugar/food)	"They use a process called photosynthesis to convert the sun's energy into glucose."	Tell me more! How is energy changed in the plant? Good start! What is the glucose molecule made from?
4c	Energy conservation	"The total energy in food chain does not change"	Nice thinking! How does energy get transformed into other types in this process?

#	Idea description	Student Example	Guidance
			Interesting! How does energy get transferred in this process?
4d	Energy is lost across the food chain	"energy travels up the food chain, or trophic pyramid, until it is all lost"	Tell me more! How does the energy get from the producers to the animals?
		"When another animal eats that animal, the energy gets transferred, but it is also less than the original energy from the plant."	Interesting! Where does the decreased energy go?
6a	Plant stores energy (in glucose)	"Plants store energy from the sun"	Great! What happens to the glucose after the animal eats the plant?
		"Plants store energy from the sun in the glucose molecules"	Nice thinking! How is the stored energy used by animals?
6b	Plants use energy/glucose/food to grow (if idea with photosynthesis, this combo is mechanistic understanding of cell resp)	"the sun helps the planters do photosynthesis and plants then use part of the glucose they made to grow bigger"	Interesting ideas! How do animals use the glucose stored in the plant?
			Good progress! How does the energy transfer from plants to animals when the animal eat plants?
6c	Animals use energy to grow/repair/hunt or animals store energy	"Animals use the energy to move, run, and hunt."	Nice thinking! What do animals do with the glucose in the plant to help them get the energy?
			Great! How does the animal make the energy usable for growth?
9	Animals take in glucose / energy (Vague without any justifications or wrong justification)	"The animal just changes the glucose"	Tell me more! How does the animal use the energy from glucose?
		"hello, animals breath in glucose to use it for energy right?"	Good progress! What happens when animals get glucose from plants?
9a	Animals use cellular respiration to release energy (or describe process or details of cellular respiration)	"the energy from the sun helps animals have enough energy to perform cellular respiration"	Nice thinking! Where does the energy that the animal can use come from?
		"Animals can make use of the sugars provided by the plants in their own cellular energy factories, the mitochondria."	Great! How do plants store energy so it can be used by animals?
11	Energy from the sun helps animals survive (vague statement depicting relationship between sun and animals)	"The sun helps animals to survive"	Interesting! How does the energy from the sun get used by the animal?
		"Animals get the energy from the sun "	Good start! How does the solar energy from the sun help animals live?
11a	Animals use sun's energy to keep warm or get vitamins	"The sunlight can keep animals warm"	Interesting ideas! How do animals use the energy from the sun?
		"Animal get vitamin D from the sun to strengthen their bone"	Good progress! How do animals get energy differently from plants?

#	**Idea description**	**Student Example**	**Guidance**
11b	Too much sunlight can be harmful (e.g. skin cancer, overheating, dehydration)	"i think too much sun will kill the animals or tired them out"	Too much sun can be harmful. What are the positive impacts of energy from the sun?
			Interesting! How can the solar energy from the sun be helpful for animals?
12	Animals eat plants/animals to survive or for food (NOT about energy)	"The plant goes through the animal's digestive system, and that absorbs the nutrients and energy an animal needs to survive."	Interesting ideas! What do animals do with the glucose they get from plants?
			Good progress! What happens to the energy after the animal eats the plant?
12b	Energy transfer: animals eat plants / animals for food / energy	"the energy of the sun helps the plants grow, animals like cows eat grass (a plant) to also get energy."	Tell me more! How does the energy from the plant get used inside the animal?
		"energy from the sun grows plants and animals eat those plants and turn their glucose into energy."	Great progress! What process does the animal do to make the energy usable?
13a	Animals get energy from sleeping/self recovery	"well they eat and get energy by sleeping"	I agree that animals need sleep! What other types of energy can animals get from the sun?
			I agree that animals get energy from sleeping and eating. Where does the energy in their food come from?

Appendix Table 3. Descriptive Statistics for Students with One Initial Idea: Comparison of Eliciting vs. Distinguishing Guidance (Iterations 3 and 4)

Guidance	Language	N	New ideas added		New Normative ideas added		% no new ideas
			Mean	SD	Mean	SD	
distinguish	Monolingual	133	0.77	0.88	0.14	0.41	88
distinguish	Multilingual	139	0.58	0.83	0.1	0.37	92.1
elicit	Monolingual	143	0.72	0.75	0.15	0.39	86.7
elicit	Multilingual	124	0.78	0.89	0.18	0.54	87.9

Appendix Table 4. Descriptive Statistics for Students with Two or More Ideas at Round 2: Comparison of Prompting with Least promising idea first vs. Most promising idea first (Iterations 3 and 4)

Prompt	Language	N	New ideas added		New Normative ideas added		% no new ideas
			Mean	SD	Mean	SD	
Least	Monolingual	70	0.79	1.02	0.4	0.69	70
Least	Multilingual	33	0.58	0.71	0.24	0.5	78.8
Most	Monolingual	71	0.59	0.65	0.24	0.46	77.5
Most	Multilingual	55	0.65	0.7	0.33	0.51	69.1

Appendix Table 5. Revised Idea Rubric and adaptive guidance for the Food Justice AI Dialog at Iteration 2 (Green colored ideas are normative ideas; blue colored ideas are vague/broad; orange colored ideas are intuitive/inaccurate)

#	**Idea description**	**Student examples**	**Guidance**
	Nonscorable	*I don't know.*	Let's think together! How does it matter if some people live closer than others to places that sell fresh fruits and vegetables? Some grocery stores have more fresh fruits and vegetables than others. How does this impact people's access to fresh foods?
1	All people have equal access	*I think all people should have access to fresh fruits and vegetables because its something you could easily grow.*	Why might it be harder for some people to get fresh fruits and vegetables, even if they are easy to grow? What could make it harder or easier to get fresh fruits and vegetables?
1b	Not all people have equal access to fresh fruits and vegetables (claim without evidence)	*Not everyone has equal access. It's obvious.*	Can you give an example of unequal access to fruits and vegetables? What are some things that contribute to some people having less access?

#	**Idea description**	**Student examples**	**Guidance**
3	Vague location or Country/city specific (larger scale), like Africa, town, Oakland, etc	*It really depends on where you live*	Interesting! Why do people who live in some places have more access to fruits and vegetables than others? Great! Tell me about a neighborhood that has less access to fresh fruits and vegetables than another one.
4	Description of availability of gardens (w/o mentioning anything that causes it), like no place to grow food	*another possible scenario could that <u>there isn't a garden that produces fresh fruits and vegetables at your neighborhood</u>* *No because some people might not have the space to grow fruit*	Good point! What determines whether a neighborhood gets gardens, farmers markets, or grocery stores? Good point! Why do you think some neighborhoods don't have gardens or places that produce fresh food while others do?
6	Description of availability of fast food (without mentioning anything that causes it)	*some communities have more fast food like burger king and 7-11.*	Interesting ideas! Why might fast food in the neighborhood make access to fresh food less likely? Nice thinking! How do fast food restaurants prevent stores with fresh fruits and vegetables in some neighborhoods?
7	Description of availability of grocery stores/supermarket/farmers market (without mentioning anything that causes it)	*I said no because not in all areas they have supermarkets and farmers markets. My neighborhood only has one Safeway.*	Interesting ideas! Why do you think supermarket chains open fewer grocery stores in some neighborhoods than others? Great! What might cause there to be more grocery stores in certain neighborhoods?
8	Lack of natural resources to grow food: sunlight, water, soil, or other resources to grow plants.	*People don't get access to fresh fruits and vegetables because some places might not have rain or enough sun light.*	I see that you are talking about growing fresh food. What if someone wants to buy fresh food? I agree! What other factors do you think make it harder for some neighborhoods to buy fresh food?

#	Idea description	Student examples	Guidance
9	Far away from grocery/far drive (Compared to idea 7, idea 9 talks directly of distance to resources. Idea 7 is about numbers of grocery store nearby)	*Some People do not have the same access to fresh fruit or vegetables because some live far away from a nearby grocery store*	Great start! Why do you think some neighborhoods don't have as many grocery stores? Nice thinking! Why is it harder for some people to find a nearby grocery store?
10	Food desert	*some people live in food desert* *We know there are food apartheid where people of color live and they don't have proper grocery stores*	Interesting! What causes a food desert? Good idea! Can you tell me more about who is most impacted in a food desert area?
11	Entrepreneurs open fast food in food deserts	*where people don't have as much money as other places they build cheaper places to eat because they can't afford healthy food*	What policies might make it easier for fast food to open in these neighborhoods instead of grocery stores with fresh produce? How does this impact people who live in these neighborhoods?
12	12-Factors related to ppl's SES - don't have money or cannot afford - Live in a low-income area - Lack of car - low job pay -unhouses/homeless	*not all the people have the same access to fresh food because of how good the income is around the area so if you live in a low income area* *when some people doesn't have jobs also some people are homeless and not having any access to fruits and vegetables*	Yes, money is important! What are other factors preventing people from getting fresh food? Yes, cost is important! Some people who have jobs and money still don't have easy access to fresh fruits. Why?
13	Lack of transportation	*Some people dont have transportation to get there.*	Good point! Why do you think some communities have less access to good public transit than others? What policies might affect whether neighborhoods have public transit to grocery stores?

#	Idea description	Student examples	Guidance
14	systemic discrimination (like redlining)	*West Oakland is subject to a history of discrimination and red lining so people in these areas have never really had those resources. Some still don't have it today.*	Think of the plans you created in the unit. How could you take action to make changes?
15	Fresh food is expensive	*fresh fruits and vegetables are very expensive sometimes*	Yes, fresh food is expensive. How might price limit who has access to fresh food? I agree! What do you think makes fresh food more expensive in some neighborhoods?
17	Race/observed correlation between ethnicity and access	*Areas with people of color don't have as many fresh fruits.*	You mentioned people of color might have less access. What caused this? Good point! Why do you think there is less fresh food in certain neighborhoods?
18	Individual Discrimination/ White privilege	*i think that people think that they are people of color they shouldn't have fresh food and it's only for white people so they only have unhealthy food available for people of color.* *People of color can only buy frozen food from 7-11 while white people have nice safeways.*	Possible prompts: ~~Who are the "people" who hold this opinion? Do you agree with them?~~ Keep going! What caused this in the past? Why would people want to control/limit what foods you could eat? Think of the plans you created in the unit. How could you take action to make changes to address this problem?

#	Idea description	Student examples	Guidance
22	Social transformation	*people in West Oakland are trying to build more stores and popup stalls that sell healthy food*	How do you think this unequal access happened in the first place? Taking action matters! Sometimes communities need to take action because they are already feeling the effects of lack of fresh food. Why do you think that is?
23	Heath related ideas	*maybe health conditions making it hard to find fruits and vegetables that meet their health costs* *Eating fast food will make you unhealthy and fat*	Yes, food and health are closely connected! How does lack of fresh food impact health differently in various communities? Some people might eat fast food more than they would like to because that's the closest option near their home. Why do you think that happened?
Below are the ideas that the model cannot accurately detect			
2	Fresh food is personal choice	*Some people just don't like eating vegetables*	
14b	Vague discrimination	*It is all about discrimination*	
16	Fast food is cheaper than fresh produce	*Some people prefer fast food because Fast food make offers like 5 pieces of chicken for three dollars*	
20	Resources to grow plants	*Some people having no resources or knowledge on growing plants*	
22	Social transformation	*people in WO are trying to build more stores and popup stalls that sell healthy food*	How do you think this unequal access happened in the first place? Taking action matters! Sometimes communities need to take action because they are already feeling the effects of lack of fresh food. Why do you think that is?

Appendix Table 6. Knowledge Integration and Idea Development Summary

Teacher	KI and Idea changes	Pretest → Embedded test → Posttest → Delayed Posttest (Six months later)	Significant Gains
Amy	KI Score at Revision	2.65→2.87→3.07***→2.80*	Pre→Post: β=1.62***; Pre→Del: β=0.58*
	Normative ideas		
	Energy Transfer	59.7% → 74.1% → 84.8%*** → 70.4%	Pre→Post: +25%***
	Photosynthesis	13.2% → 31.0%* → 45.4% → 42.0%***	Pre→Emb: +18%*; Pre→Del: +29%***
	Animal Cellular Respiration	3.3% → 14.1% → 41.2%*** → 21.0%	Emb→Post: +27%***; Pre→Post: +38%*** (after drawing)
	Photosynthesis Reactants	1.5% → 10.2% → 15.2% → 8.0%	ns
	Photosynthesis Products	5.0% → 13.5% → 23.3% → 18.6%	ns
	Animals use energy to grow/hunt	9.0% → 13.8% → 25.8% → 27.8%	ns
	Intuitive and Precursor Ideas		
	Sun helps plants grow	58.2% → 37.9% → 42.4% → 37.0%	ns
	Sun helps animals survive	35.8% → 39.7% → 40.9% → 46.3%	ns
	Animals eat plants for food and nutrition, not energy	47.8% → 48.3% → 40.9% → 44.4%	ns
	Animals directly use sun to keep warm or get vitamins	25.4% → 21.9% → 20.2% → 23.4%	ns
	Other ideas (3a,4a,4b,4c,4d,6a,6b, 9,13a)	<10% across all timepoints	ns

Ben	KI Score at Revision	2.32→2.63→2.69*→2.48	Pre→Post: β=0.35*; Pre→Del: β=0.08
	Normative ideas		
	Energy Transfer	36.5% → 40.0% → 50.2% → 43.7%	ns
	Photosynthesis	9.2% → 29.8% → 34.7%* → 26.4%	Pre→Post: +26%*
	Animal Cellular Respiration	1.6% → 7.7% → 11.7% → 7.2%	ns
	Photosynthesis Reactants	1.0% → 29.8%* → 29.8% → 20.0%	Pre→Emb: +29%* (after labs)
	Photosynthesis Products	1.6% → 21.5%* → 17.6% → 15.0%	Pre→Emb: +20%* (after labs)
	Animals use energy to grow/hunt	7.7% → 16.0% → 15.8% → 13.2%	ns
	Intuitive and Precursor Ideas		
	Sun helps plants grow	28.8% → 38.0% → 26.3% → 28.3%	ns
	Sun helps animals survive	23.1% → 18.0% → 28.9% →15.1%	ns
	Animals eat plants for food and nutrition, not energy	36.5% → 32.0% → 31.6% → 26.4%	ns
	Animals directly use sun to keep warm or get vitamins	13.3% → 10.2% → 9.5% → 11.3%	ns
	Other ideas (3a,4a,4b,4c,4d,6a,6b, 9,13a)	<10% across all timepoints	ns

Note: Percentages show proportion of students expressing each idea in revised explanations after the dialog. KI scores are mean revised scores (after dialog support; range 1-5). Statistical significance from GEE binomial tests for ideas and CLMM for KI scores: ***p<.001, *p<.05.