International Journal of Artificial Intelligence in Education (2019) 29:283–316
https://doi.org/10.1007/s40593-019-00178-y

ARTICLE

Problematizing Helps! A Classroom Study of Computer-Based Guidance for Invention Activities

Catherine C. Chase[1] · Helena Connolly[1] · Marianna Lamnina[1] · Vincent Aleven[2]

Published online: 9 April 2019

Abstract

A successful instructional method is to engage learners with exploratory problem-solving before providing explanations of the canonical solutions and foundational concepts. A key question is whether and what type of guidance will lead learners to explore more productively and how this guidance will affect subsequent learning and transfer. We investigate this question through the design and study of the *Invention Coach*, an adaptive, computer-based learning environment that problematizes students' understanding as they invent fundamental physics equations. Problematizing guidance (Reiser *Journal of the Learning Sciences, 13*(3), 273–304, 2004), which encourages learners to grapple with domain complexity, is well-suited to the goals of Invention. However, there are few examples of technology-based learning environments that were explicitly designed to problematize and scant research on their efficacy. In an experimental study, 199 middle schoolers worked with either *motivational, task + motivational, or problematizing + task + motivational* guidance versions of the Coach while inventing. Students who engaged with the problematizing Coach were better able to transfer their knowledge to novel domains in the short term, and their transfer gains were comparable to those provoked by human tutors. While students in the problematizing condition were less likely to invent the correct solutions, they engaged in more targeted and efficient exploration of the solution space and were less likely to report experiences of difficulty. Findings suggest that problematizing guidance has the potential to effectively support exploratory problem-solving, when the goal is to facilitate productive exploration and transfer from subsequent instruction. The work also has implications for the design of problematizing guidance.

Keywords Invention · Exploratory learning environments · Scaffolding · STEM learning · Adaptive technologies · Productive failure

✉ Catherine C. Chase
chase@tc.columbia.edu

Extended author information available on the last page of the article

Introduction

In many traditional forms of science and math instruction, teachers explain core concepts and formula, then ask students to apply them in practice problems. In contrast, *explore-then-explain* (ETE) instruction flips the traditional script on its head by first engaging learners in exploratory problem-solving before teachers explain the canonical solution and related concepts. While students often fail to successfully solve the problem during the explore phase, their exploration is often productive. While exploring, students may investigate a variety of problem solutions, developing well-differentiated knowledge of the domain (Kapur 2008; Schwartz and Martin 2004). This, in turn, may prepare students to learn and transfer from provided explanations of the target knowledge, which often take the form of lectures or readings (Bransford and Schwartz 1999). There are several classes of ETE approaches, such as productive failure (Kapur 2008) and problem-solving prior to instruction (Loibl and Rummel 2014). In this article, we focus on an ETE approach called Invention (Schwartz and Martin 2004). During Invention activities, students attempt to generate an external representation of a deep domain structure that runs throughout a set of contrasting cases. For example, in the current study, students attempt to invent equations for physical science phenomena which contain the deep mathematical structure of ratio (e.g. density = mass/volume, speed = distance/time) before receiving a lecture on ratio structures in physics. Invention and other ETE approaches have been shown to improve conceptual learning and transfer over traditional instruction, in several studies, spanning ages from adolescence to adulthood, and in many science and math domains (Kapur 2008, 2010, 2011, 2014; Kapur and Bielaczyc 2012; Loibl et al. 2017; Schwartz and Bransford 1998; Schwartz and Martin 2004; Schwartz et al. 2011).

A key question is whether and what type of guidance during the exploration phase will lead learners to explore more productively and how this guidance will affect subsequent learning and transfer. The literature on discovery learning, which is a close cousin to exploration, has long established that discovery is more effective than direct instruction, but only when it is guided (Alfieri et al. 2011; Mayer 2004). In the guided discovery paradigm, the learner is typically guided towards discovering the correct solution. This kind of structuring guidance, which generally simplifies the task for learners, is a common form of scaffolding (Reiser 2004; Wood et al. 1976). However, others suggest that only minimal guidance during the explore phase of ETE paradigms is necessary. One study of guidance during the explore phase of ETE instruction found that students who only received affective support learned as much as students who received on-demand help from the teacher and intermittent whole-class lectures and discussions (Kapur 2011). The author suggests that “it may well be more productive to delay help even when students ask for it, and perhaps first give them an opportunity to find a way out themselves.” From this viewpoint, failures and the further exploration they provoke are valuable learning opportunities, and structuring forms of guidance which lead learners to success only reduce that exploration. Thus, teachers should limit their guidance. Others suggest that students’ exploration can be made even “more productive” by supporting students’ inquiry behaviors (Holmes et al. 2014). We also aim to enhance the productivity of students’ exploration, but through the use of *problematizing* guidance (Reiser 2004). In contrast to structuring forms of guidance which simplify the task, problematizing guidance engages learners with domain complexity.

Problematizing guidance helps learners realize that some aspect of their thinking is problematic, encouraging learners to confront and grapple with key disciplinary ideas.

Problematizing guidance is an appealing alternative to more structured forms of guidance that lead learners by the nose down a correct solution path, cutting short students' exploration of the domain space. However, it is not clear how the construct of problematizing guidance should be operationalized. There are few examples of technology-based learning environments that were explicitly designed to problematize learners' thinking, and even fewer that problematize adaptively. Moreover, there is little research on the efficacy of problematizing guidance. Finally, the research on ETE approaches has generated mixed evidence on whether and which forms of guidance are beneficial during the explore phase.

In this paper, we describe the design and implementation of problematizing guidance in the *Invention Coach*, an adaptive, computer-based learning environment designed to guide students through the exploratory phase of Invention. We then present findings from an empirical study that explores the efficacy of our full problematizing version of the Coach as compared to two minimally guided versions, which receive only motivational guidance or motivational guidance plus support for understanding task goals and constraints. We examine how these different guidance configurations affect middle school students' exploration, learning, and transfer of ratio structures in science. Note that we explicitly use the broader term "guidance" instead of "scaffolding," which implies later fading.

Structuring and Problematizing Guidance

A useful distinction in technology-based guidance is between structuring and problematizing forms of guidance (Reiser 2004). Structuring forms of guidance *simplify the task* in some way for learners, making it possible for a learner to accomplish complex tasks or solve messy problems that they could not do on their own. Two common ways to structure the task are to decompose it into component parts and to narrow learners' choices. For example, inquiry "checklists" can help learners follow the steps of the inquiry process (Linn et al. 2004). In addition, many simulations narrow the range of variables to test by providing options to vary only a few factors. Likewise, many traditional intelligent tutoring systems provide extensive structuring guidance by breaking the problem into subgoals and steps and providing correctness feedback (VanLehn 2006). In our view, most forms of guidance in technology-based learning environments structure the task in some way (Reiser 2004).

While structuring guidance makes the task easier, problematizing guidance strives to engage the learner with domain complexity. Problematizing guidance helps learners recognize that their understanding is problematic and encourages them to "encounter and grapple with" key disciplinary ideas that they would otherwise overlook (Reiser 2004). Problematizing guidance often takes a cognitive conflict approach, leading learners to bump up against key domain structures and principles which often contradict their current understanding. Problematizing guidance differs from structuring guidance in two key ways. First, structuring guidance supports *completion of the task*, while problematizing guidance supports learners' *understanding of the domain*. Second, structuring guidance *simplifies*, while problematizing guidance *complexifies*, provoking a desirable difficulty of sorts (Bjork 1994). In this way, problematizing guidance can make the task more difficult in the short term but may ultimately prove more productive for learning.

Prompting learners for explicit articulation of ideas and decisions, engaging them with the language and representations of the discipline, and surfacing knowledge gaps or disagreements are specific ways in which guidance can problematize. For example, ExplanationConstructor, a computer-based journal for supporting inquiry, requires students to classify their scientific explanations according to a type of explanatory framework, such as natural selection (Sandoval 2003). Ensuing discussions around this task uncovered students' disagreements about what constitutes a "trait" and ideas about the relationship between structure and function (Reiser 2004). By eliciting decisions that require learners to connect their work with key disciplinary ideas, a technological tool can encourage students to grapple with deep domain principles and refine their thinking. A second example comes from the Betty's Brain teachable agent software (Biswas, Leelawong, Schwartz, Vye, and The Teachable Agents Group at Vanderbilt 2005), in which learners teach by essentially programming their agents to reason through a series of links in a causal concept map. The agents then go on to answer questions in a gameshow, but first, learners have to make predictions about how their agents will answer (Chase et al. 2009). Students often make a prediction based on their own knowledge and are surprised when this conflicts with what their agent "knows" (Chase et al. 2009). This can problematize learners' understanding of how the agent reasons through links in a causal chain, which relate to underlying biological mechanisms and more general causal reasoning skills. This learning mechanic of predicting then observing can problematize a learners' understanding when the prediction conflicts with the outcome.

As may be evident from the above examples, problematizing is a large umbrella term, which can take many different forms. For instance, problematizing forms of guidance may encompass cognitive conflict (Piaget 1977) and socratic tutoring (Collins et al. 1975) approaches, which both lead learners to a contradiction in their thinking, provoking deeper engagement with domain complexity.

Another caveat is that many forms of guidance simultaneously structure and problematize (Reiser 2004). For instance, ExplanationConstructor also structures by providing learners with a limited set of explanatory frameworks to choose from, rather than asking them to generate their own. Likewise, metacognitive prompts structure the process of monitoring one's understanding whilst helping learners notice problematic gaps in their knowledge. Thus, it may be challenging to design guidance that effectively problematizes without also providing some form of structure.

Problematizing guidance is well-suited to Invention and other exploratory problem-solving tasks. This is because problematizing guidance and Invention tasks are designed to provoke similar learning mechanisms. In a review of several studies of ETE instruction, Loibl et al. (2017) identified three learning mechanisms that can be provoked during the exploration phases of ETE: activating prior knowledge, revealing knowledge gaps, and recognizing deep domain features. These overlap nicely with the goals of problematizing, which are to realize that some aspect of one's thinking is problematic and encounter and grapple with key disciplinary ideas.

A potential danger of problematizing guidance is that it may be frustrating for learners. A goal of problematizing guidance is to make learners realize that their undertanding is problematic. This could lead learners to flounder and experience negative affect. In addition, problematizing guidance might be ineffective for low performers, who may experience the most floundering and frustration. Thus, we tested whether problematizing guidance had differential effects for learners with different levels of prior knowledge. We also conducted exploratory analyses of learners' experience of problematizing guidance.

Designing to Problematize

Few scholars have explicitly designed problematizing guidance for technology-based learning environments (Efstathiou et al. 2018; Hicks and Doolittle 2008; Molenaar et al. 2011). In addition, problematizing guidance has been operationalized in a variety of different ways, including analysis questions (Hicks and Doolittle 2008), graphic organizers (Chen et al. 2011), reflection prompts (Feng and Chen 2014), feedback on one's mistakes (Efstathiou et al. 2018), and self-regulatory prompts (Molenaar et al. 2011). This is perhaps because Reiser's framework stops short of providing explicit guidelines for designing software supports for problematizing. It is possible that problematizing guidance is an open-textured concept – a concept whose meaning becomes clear, and may even change, as it is applied to specific problems (Lynch et al. 2009). Thus, it is necessary to further explore how problematizing guidance might be optimally designed. We have created a set of design guidelines for our brand of problematizing guidance, as it pertains to exploratory problem-solving.

Moreover, there is scant research on the efficacy of technology-based problematizing guidance. Efstathiou et al. (2018) found that software with combined problematizing and structuring guidance was more effective at enhancing students' experimental design skills than a paper-and-pencil version that only structured the task. Molenaar et al. (2011) pitted structuring and problematizing metacognitive prompts against one another and found that problematizing prompts led to greater learning of metacognitive knowledge and greater transfer of domain knowledge. Hicks and Doolittle (2008) implemented both structuring and problematizing guidance into a multimedia tutorial on historical inquiry and found that learners exposed to this instruction made strides in retention and application of the taught inquiry strategy. While these studies demonstrate positive effects for problematizing guidance, considerably more research is needed to explore the efficacy of problematizing guidance in its various forms.

Interestingly, the problematizing framework has rarely been instantiated in an intelligent learning environment that is highly adaptive (but see Efstathiou et al. 2018, who problematized learners' mistakes). Adaptive technologies can provide highly customized, individualized feedback and guidance that adapts to a student's solution path, current knowledge state, or other individual characteristics (Aleven et al. 2016b; Shute and Zapata-Rivera 2012). Adaptive problematizing may be a particularly effective form of guidance because it can be customized to target the precise aspect of the learners' understanding that is problematic. However, many adaptive, intelligent learning environments, provide fairly explicit, structuring forms of guidance (VanLehn 2006). Few of them explicitly problematize.

Generativity & Exploration

Finally, an issue that neither Reiser's (2004) framework nor existing empirical studies of problematizing technologies address is how problematizing guidance influences learners' generative and exploratory processes. While highly structured forms of guidance can lead learners down a specific solution path, cutting exploration of other solution paths short, problematizing guidance instead guides learners towards deep structures and concepts that underlie a domain, without necessarily specifying a solution path.

We view the exploration of the problem/domain space through generation of multiple problem solutions as an important goal of ETE instruction. In general, constructive learning activities that encourage learners to generate their own inferences and ideas tend to be productive for learning (Chi 2009). Also, in much of the research on ETE instruction, generating the right solution is less important than generating many solutions. Several ETE studies have found a positive relationship between the diversity of generated solutions and learning outcomes, suggesting that broad exploration begets learning (Kapur, 2012; Kapur and Bielaczyc 2012; Wiedmann et al. 2012). However, ETE studies conflict on how guidance influences students' exploration. One study found that extensive guidance from a teacher during exploratory problem-solving reduced the number of solutions student generated compared to a minimally guided condition (Kapur 2011). We suspect that this guidance may have overly structured the task for learners, cutting off their exploration. For instance, another study found that guidance in the form of meta-cognitive reflection prompts pushed students to generate a greater diversity of solutions compared to an unguided condition (Roll et al. 2012). In this paper, we investigate how our problematizing guidance affects not just learning but also learner's generativity and exploration of the problem space, relative to two minimally guided conditions. We do this by examining the quantity, diversity, and quality of solutions generated, and also the rate at which solutions were generated.

The Current Study

In the current study, we explore how our designed problematizing guidance impacts students' exploration, learning, and transfer. We also test how our problematizing guidance compares to two more minimal forms of guidance. In the *motivational guidance* condition (M), students received motivational messages encouraging them to persist. This is similar to the successful minimally guided condition in Kapur (2011), in which students were only encouraged to "persist and think of solutions for themselves." In the *task + motivational guidance* condition (TM), learners received motivational support plus reminders about task goals and constraints. Since the reminders merely reiterate the given goals and constraints of the task, this guidance served to clarify the nature of the task. This condition was added because pilot data indicated that students often had difficulty understanding the Invention task itself. Thus, it is possible that merely clarifying the nature of the task with some motivational support is all the guidance learners need, leaving them the space to explore expansively. Our third condition tests whether the addition of problematizing guidance provides benefits for learners. In the *problematizing + task + motivational* guidance condition (PTM), students received motivational messages, task reminders, and problematizing activities, which were designed to adaptively problematize their understanding of ratio structures in science. Conditions were designed to be incremental, testing the additive effect of each form of guidance.

In comparing these conditions, we tested the following hypotheses and exploratory questions:

- Relative to comparison conditions, students in the PTM condition should develop deeper learning (learning and application of concepts) that is more likely to transfer to novel situations. These differences should hold up on a time delay.
- How will PTM, TM, and M forms of guidance impact Invention task performance (i.e. invention of the correct solution)? We did not have a strong hypothesis going into the study. On the one hand, problematizing guidance pushes learners towards domain understanding rather than task success. One the other hand, deep understanding and task success are often linked.
- Relative to comparison conditions, students in the PTM condition should engage in more productive exploration of the problem space as measured by breadth, quality, and rate of solution generation. However, the quantity of generated solutions might be lower in the PTM condition since problematizing guidance could take time away from generating.
- What kinds of instructional experiences will PTM, TM, and M versions of the Invention Coach system create for students? We examined how often and how long students used various components of system guidance. We also explored spontaneous comments students made during reflection on the Invention tasks.

This study aims to make several contributions to the literature. First, our study adds a data point to the debate about what form of guidance (if any) can effectively support learners during the explore phase of ETE instruction. Second, we test the efficacy of problematizing guidance, with a focus on how it impacts students' generativity and exploration, which has rarely been done. Third, we provide a novel example of an adaptive technology that was explicitly designed to problematize learners' thinking. Finally, we generate a set of broad guidelines for designing problematizing support for exploratory problem-solving, to articulate our general problematizing approach.

Invention Coach Design

In developing technology-based guidance for Invention activities, we set out to explore one small corner of the design space of problematizing guidance. It is an incredibly large design space, as evidenced by the large variety of problematizing examples given in Reiser's (2004) article, and the variety of ways in which problematizing scaffolds have been operationalized (Chen, Looi, & Wen, 2011; Efstathiou et al. 2018; Feng and Chen 2014; Hicks and Doolittle 2008; Molenaar et al. 2011). In this section, we outline the design of the "full" Invention Coach system, which includes problematizing guidance, task reminders, and motivational messages. Note that we provide an intuitive description of the system here, rather than a highly technical one, which is beyond the scope of this paper and has been described elsewhere (Aleven et al. 2017). Likewise, we present the broad design guidelines we followed, which we hope will prove applicable for other designers of problematizing guidance for ETE instruction.

Problematizing Design Guidelines

In the process of designing and building the Invention Coach, we generated our own interpretation of problematizing guidance as it pertains to exploratory problem-solving, and distilled these into a set of three problematizing design guidelines.

First, *problematizing activities should draw on* instructional strategies that reveal knowledge gaps and encourage learners to notice deep features of the domain. In our design we employed contrast and explain strategies. Prrompting learners to contrast cases can help learners notice overlooked features (Bransford et al. 1989). Explanation-based activities can activate prior knowledge and reveal knowledge gaps (Chi et al. 1994). They can also be designed to engage learners with disciplinary terms. These instructional strategies are well-aligned with the goal of problematizing because they support learners in realizing that their current understanding is problematic and confronting key disciplinary ideas. Contrast and explain strategies have also been used successfully in the Invention Support Environment, a non-adaptive computer-based system (described in Holmes et al. 2014).

Second, *guidance should explicitly refrain from direct "telling,"* such as explaining to learners what to know or do. Avoiding direct telling forms of guidance maintains the open nature of the task, giving learners space to explore. Also, our prior study revealed that the more teachers posed generative questions and the less they explained while guiding an Invention activity, the more students transferred their knowledge to novel tasks (Chase, Marks, Bernett, Bradley, & Aleven, 2015). Therefore, the Coach never tells students the correct solution, exactly what is wrong with their solution, nor what to do next.

Our third design guide was to *model interactions after human teachers*. This can make the guidance feel more naturalistic while revealing appropriate instructional strategies. Many of our ideas drew on activities we observed experienced teachers devising for students as they guided them through paper-and-pencil Invention activities in an earlier study (Chase et al. 2015). These activities included inviting learners to rank cases, contrast specific pairs of cases, and explain their solution methods (which nicely overlap with the instructional strategies discussed in our second guideline).

Invention Activities

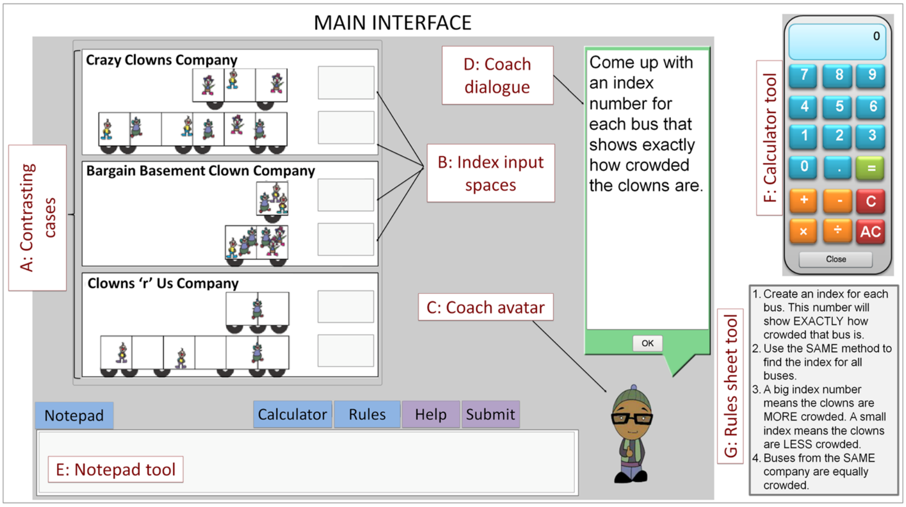
Before describing the Invention Coach design, we must first describe Invention activities in greater detail. Figure 1A shows one of the Invention activities guided by the Coach (adapted from Schwartz et al. 2011). The objective is to design a numerical index of "clown crowdedness" that expresses how crowded the clowns are in each bus (the correct solution is # clowns / # bus compartments). Students receive a few constraints: buses within the same company are equally crowded, a larger index number means a bus is more crowded, and use the same method to find the index index for each bus. In this activity, crowdedness is a proxy for density, which is introduced during a subsequent lecture. Though they don't know it, students are essentially inventing the formula for density (density = mass/volume).

Students are often unsuccessful in creating the target formula, but the process of Invention prepares them to learn more from subsequent direct instruction. Although the Invention activities have clear right answers, for the students, these activities are highly unusual and ill-defined; typically, students have never seen this kind of task before.

Invention often employs "contrasting cases," which systematically vary on deep features of a domain (Bransford et al. 1989). For example, in Fig. 1A, the size of the bus and number of clowns vary across cases such that several pairs of cases hold one feature constant while varying the other. Comparing and contrasting cases may help students notice critical features of crowdedness and identify the common ratio structure (clowns: compartments) that runs throughout all cases (Schwartz et al. 2011). However, without prompts to compare and generalize across cases, students may focus on surface features in the cases (Shemwell et al. 2015; Roll et al. 2012).

System Overview

Figure 1 shows the main screen where students work on the Invention activities, with guidance from the Coach. The students design and submit their invented solutions, where a solution is a set or partial set of index numbers for each of the cases in Fig. 1A. The Coach is represented by an avatar image (students can choose from several possible Coach characters), shown in Fig. 1C. Each time a student submits a solution, the Coach gives adaptive guidance that responds to the current state of the student's invention, taking into account the history of the interactions between the Coach and student on the given Invention task. Below we describe the problematizing activities, task reminders, motivational messages, and tools that comprise the full problematizing version of the Invention Coach system.



Come up with an index number for each bus that shows exactly how crowded the clowns are.

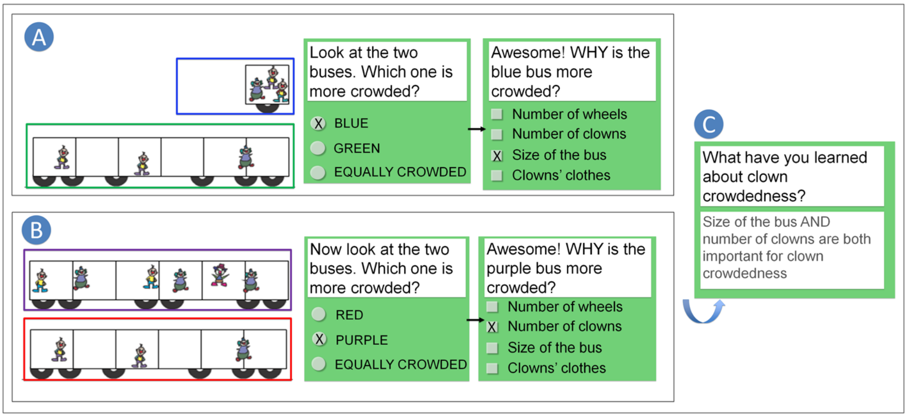
1. Create an index for each bus. This number will show EXACTLY how crowded that bus is.
2. Use the SAME method to find the index for all buses.
3. A big index number means the clowns are MORE crowded. A small index means the clowns are LESS crowded.
4. Buses from the SAME company are equally crowded.

Fig. 1 Main interface of the Invention Coach and tools. Lettered boxes (A, B, C, D, E, F, G) mark annotations of key system components

Problematizing Activities

The Coach often responds to a submitted solution with one of several interactive activities for problematizing student understanding. In these activities, the Coach engages the student in an elaborate dialogue, asking the learner to make judgments and articulate their ideas. Behind the scenes, these dialogues have elaborate branching structures. Describing them in detail, however, is beyond the scope of this paper. In descriptions of each activity below, we highlight only a small portion of the interactions between Coach and student.

Feature Contrast Students often first attempt a basic "counting" solution type, using the number of clowns in each bus as a measure of crowdedness, and neglect to consider the size of the bus. The feature contrast activity (in Fig. 2) is meant to problematize this simplistic interpretation of crowdedness. Learners are asked to contrast specific pairs of cases that differ on a key feature while holding other variables constant. It then asks the learner which of these two cases is more crowded and why. In this 'why' prompt, learners must identify a specific feature or pair of features that make one bus more crowded than another. If students identify the wrong feature(s), the system confronts them with two new cases that hold the selected feature constant while varying crowdedness and asks them again which features matter for crowdedness. The point of these contrasts is to problematize learners' notion of the broader concept (e.g. crowdedness/density), by helping them notice a deep feature of the domain their solution has overlooked. After going through contrasts relevant to each feature (Fig 2A and B), students are asked to put together what they learned from the contrasts into a singular explanation (Fig. 2c). This should help learners to notice the importance of both deep features of crowdedness (clowns and space) and consider their relation when inventing their next solution. The feature contrast activity is called when a student's solution ignores a key deep feature.



Look at the two buses. Which one is more crowded?

☒ BLUE
○ GREEN
○ EQUALLY CROWDED

Awesome! WHY is the blue bus more crowded?

□ Number of wheels
□ Number of clowns
☒ Size of the bus
□ Clowns' clothes

Now look at the two buses. Which one is more crowded?

○ RED
☒ PURPLE
○ EQUALLY CROWDED

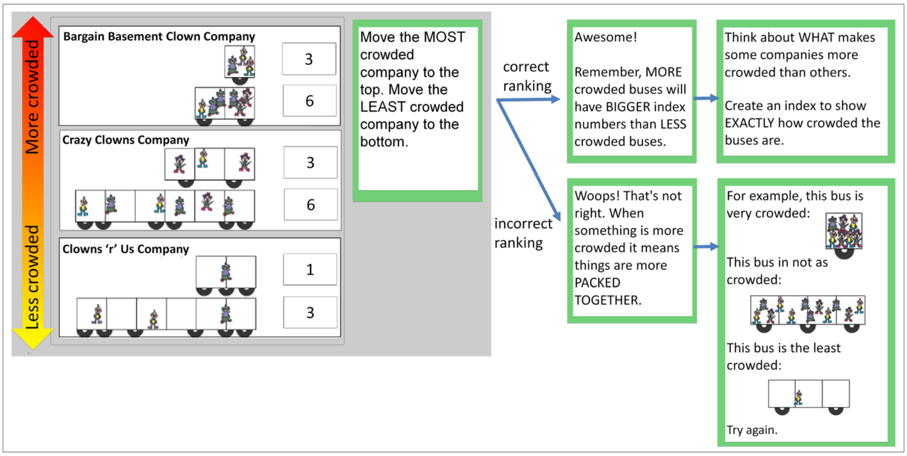
Awesome! WHY is the purple bus more crowded?

□ Number of wheels
☒ Number of clowns
□ Size of the bus
□ Clowns' clothes

What have you learned about clown crowdedness?

Size of the bus AND number of clowns are both important for clown crowdedness

Fig. 2 Feature contrast activity



Move the MOST crowded company to the top. Move the LEAST crowded company to the bottom.

Awesome!

Remember, MORE crowded buses will have BIGGER index numbers than LESS crowded buses.

Think about WHAT makes some companies more crowded than others.

Create an index to show EXACTLY how crowded the buses are.

Woops! That's not right. When something is more crowded it means things are more PACKED TOGETHER.

For example, this bus is very crowded:

This bus in not as crowded:

This bus is the least crowded:

Try again.

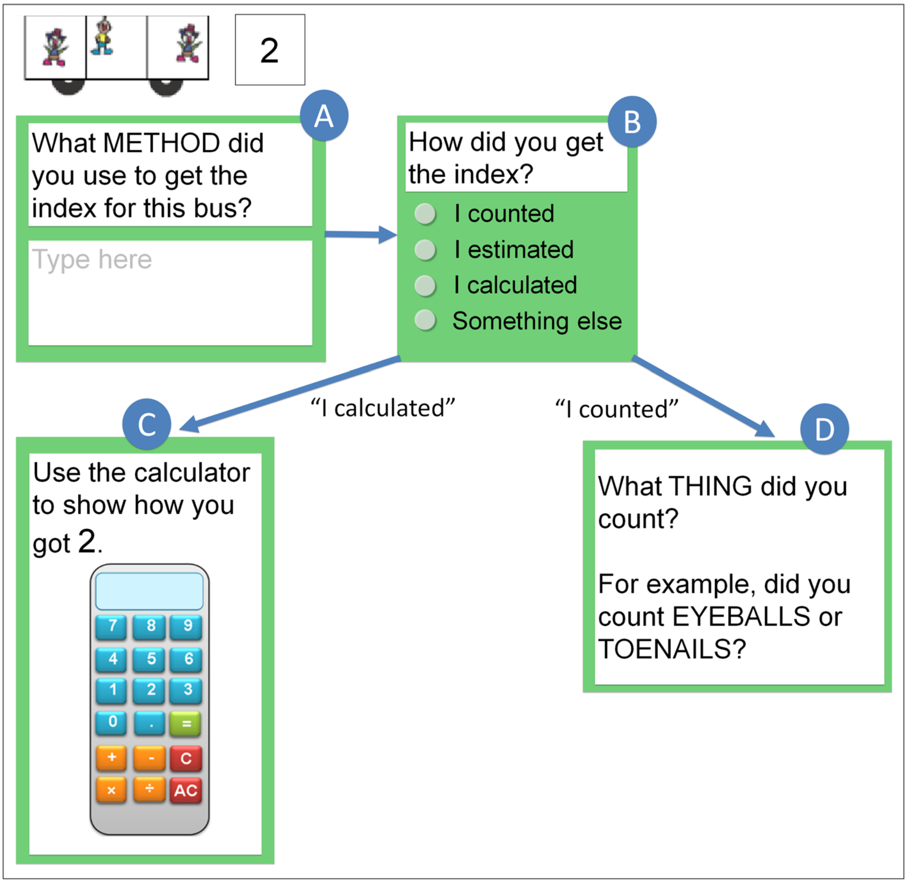
Fig. 3 Ranking activity

Ranking The ranking activity prompts students to rank the companies from least to most crowded (Fig. 3). If learners rank incorrectly, the Coach tells them that their ranking is not right, gives them the definition of the concept ("When something is more crowded it means things are more packed together"), and gives them up to three more attempts to rank the cases. When students complete the ranking activity, which almost all students do successfully, the Coach reminds them that their index numbers should rank the cases correctly.

The ranking activity is called when students' ranking does not match their index numbers. Here the ranking activity has the potential to problematize students' understanding by revealing this inconsistency to the learner. Most students have an intuitive notion of crowdedness and can distinguish between more and less crowded buses. They are often surprised when their calculated index numbers don't match their intuitive ranking. For example, in Fig. 3, the student has dragged and dropped the cases so that they are correctly ordered from most to least crowded, but their index numbers (demonstrating a single-feature counting clowns solution) do not corroborate their ranking. The ranking activity is also called when students submit "unclassifiable" solutions for which it is hard to discern a consistent method of calculating index numbers across cases. These often constitute "random guesses," and the ranking activity signals the need to think more systematically about their index numbers. In this way, the ranking activity can push the learner to evaluate the accuracy of her invented solution and may surface problems with the learner's reasoning.

Tell-Me-How The Tell-Me-How activity (Fig. 4) aims to problematize students' understanding by asking them to articulate how they arrived at their answers using disciplinary representations (e.g. numbers and operations). The Coach's prompts are meant to encourage students to reflect on their solution process, engage in precise reasoning around disciplinary ideas, and surface knowledge gaps. Tell-Me-How is called frequently, throughout an Invention session, and focuses on one particular case (bus). Students are asked to explain their invented method in a write-in response, then select whether they "counted," "estimated," or "calculated" their answer. Students are then asked to map the numbers in their solution method to referents in the cases by labelling what each number represents. Through the mere process of articulating their solutions in disciplinary terms, designating precise calculations and referents for numbers, students often come to reveal problems with their solution methods and think more deeply about mathematical ideas. For example, the student in Fig. 4 has been asked to explain how she came up with the index number 2 for one of the buses. While not shown in the figure, this student has generated what we call an estimating solution type, in which she has designated the crowdedness of the buses to be 1, 2, and 3, based on their relative ranking. The student selected the option "I calculated" to describe her method, but then realized she could not reproduce her solution using mathematical operations on the calculator and needed to develop a more precise mathematical understanding. These aspects of the activity problematize by guiding students to engage deeply with mathematical ideas they often gloss over.

In addition to problematizing, all three activities add a layer of structure by focusing the student on a specific subgoal of the task (e.g. comparing and contrasting cases), or



What METHOD did you use to get the index for this bus?

How did you get the index?

Use the calculator to show how you got 2.

What THING did you count?

For example, did you count EYEBALLS or TOENAILS?

Fig. 4 Tell-me-how activity

providing a constrained set of options to students (e.g. focusing students on two cases in the feature contrast activity). In general, we found it challenging to problematize without simultaneously providing more structure, one of the design tensions that Reiser acknowledges (Reiser 2004). However, the main goal of each activity was for learners to wrestle with key disciplinary ideas, and the additional structure sometimes guided them in this.

Task Reminders

The Coach also provides guidance in the form of brief "task reminders." These reminders are meant to help learners understand the task, since most students have never encountered an Invention task and their first reaction is often confusion. Reminders are brief messages that point out a specific goal/constraint the learners' solution is violating. For instance, if students appear to be using different invention methods across cases, the Coach will remind the student that "you have to use the exact same method to find the index for all of the buses." Or if a student's index numbers do not correctly rank the cases from most to least crowded, the Coach will say: "A big index number means the clowns are more crowded. A small index number means the clowns are less crowded." There were essentially three kinds of reminders tied to problem constraints, which pushed the learner to create a solution that is (1) general enough to work for all cases, (2) precise, and (3) accurate. Reminders were given after students submitted their solution or clicked "help."

We do not view these reminders as problematizing forms of guidance because they do not explicitly encourage learners to connect with disciplinary ideas. Whereas problematizing activities guide learners to encounter the deep features of the content (using feature contrast to highlight features, ranking to make relevant prior knowledge salient, or tell-me-how to encourage explanations in disciplinary terms), reminders focus solely on the problem solution. Moreover, since the goals and constraints of the task are given as part of the initial Invention task instructions, these reminders are not providing new information. They merely reiterate the given goals and constraints adaptively, to point out when one of them is violated.

Motivational Messages

Every time students click "help" or submit a solution, they receive a brief motivational message. Messages encourage effort and emphasize the progress students are making, such as "Your brain is getting a workout! Keep it up!" or "You're making good progress!" These messages are meant to alleviate students' frustration and encourage persistence during the very challenging task of Invention.

Tools

In addition, learners have access to several self-help tools (Fig. 1E, F, G), which they can open at any time, including a calculator, a notepad where they can jot notes, and a rules sheet that lists all task goals and constraints.

Adaptivity

We designed the Invention Coach to be highly adaptive to the student's approach to Invention, because to truly problematize, the system must interpret the student's solution and respond with an appropriate form of guidance that helps learners realize the flaw in their understanding that is reflected in their solution. Secondarily, the Coach keeps track of guidance that has been tried before, so as to avoid repetition. The Coach's adaptive algorithm is described in our previous paper (Aleven et al. 2017). Here we give a brief overview. We note that the algorithm assumes a ratio structure for the correct invention.

Each time a student submits a solution or clicks on the help button, the Coach interprets the solution and responds with a relevant form of guidance. The Coach interprets the solution by classifying it into one of five broad categories, based on the index numbers the student entered, and when available, explanations from the Tell-Me-How module. The categories were generated based on our pilot studies. They represent student attempts at inventing a ratio structure, with increasing sophistication. *Single-feature* solutions count one of the deep features of the domain (e.g. clowns or bus compartments in the crowdedness task). *Two-feature* solutions consider both deep features in creating their index numbers, but do not relate them mathematically (e.g. most of these solutions provide a rating or ranking of the cases along both dimensions). *Mathematical two-feature* solutions relate the two deep features using an incorrect mathematical operation (addition, subtraction, multiplication, etc.). *Ratio-based* solutions relate the two deep features of the domain using any possible variant of a ratio structure (e.g. inverse of correct ratio, correct ratio doubled, etc.). Finally, *unclassifiable* solutions do not fall into any of the above categories and often represent seemingly random guesses. To be placed into one of these categories, more than half the index numbers submitted need to be consistent with the category.

Next, based on solution category, the Coach selects a form of guidance to confront the learner's most egregious problem in their thinking or gap in their knowledge. For each solution category, there are several task reminders and problematizing activities that the system can provide, ordered in terms of how specifically they address the student's solution. If the student continues to submit the same type of solution,[1] the Coach cycles through these options, alternating between reminders and activities. When the student submits a different solution type, the system will start running through the sequence of reminders and problematizing activities associated with that new solution type. For example, if a student submits a *Single-feature* solution (e.g., counting the number of clowns in a bus), the Coach might initiate a Feature Contrast activity, selecting a pair of contrasting cases that may help the learner realize the importance of a second, overlooked feature of the domain. Alternatively, the Coach could give a reminder that "a big index number means the clowns are more crowded" (this constraint is violated when index numbers are based on single features). Similarly, if a student has generated a mathematical 2-feature solution that uses the wrong

[1] Note that in the TM condition, task reminders were adaptively selected using the same algorithm described here, but problematizing activities were removed from the set of guidance options. Thus, adaptive guidance for the TM condition consisted of adaptively selected task reminders only. Also the PTM condition had additional adaptivity within the problematizing activities, in which the Coach's selection of prompts depend on the student's input during the activity.

mathematical operation, the guidance can prompt them either with a reminder that “buses from the same company should be equally crowded” (this constraint is violated when index numbers are computed with wrong math) or invoke the Tell-Me-How activity, so the student can reflect on how she calculated the index.

This description omits many details that we do not have the space to describe. The Coach’s classification strategy employs a detailed set of conditions for the five main categories of student input, as well as 21 subcategories that make up variations within the five main categories. The interactive activities Feature Contrast and Tell-Me-How each have an elaborate branching structure specifying how the Coach’s next prompt depends on the student’s answer to the previous prompt, with many different paths through the dialogues. As a measure of its complexity, the Coach is made up of 140 production rules and 150 functions. The Coach is implemented in CTAT (Aleven et al. 2016a), with the adaptive algorithm (including the classification method and the method for selecting guidance based on the classification) implemented as production rules.

Methods

Participants

The final sample[2] contained 199 students from 9 seventh- and eighth-grade classes who participated in all days of the study and made use of the guidance features in the Invention Coach system. Overall, 92 seventh-graders and 107 eighth-graders participated. Student participants hailed from a low-to-average performing public middle school in New Jersey whose population was 96% Hispanic, 56% male, and low socio-economic status (87% of students receive free or reduced-price lunch). The study was conducted during regular science class periods. Condition was randomized at the student level. Students within the same class were randomly assigned to one of three different conditions: PTM ($n = 68$), TM ($n = 70$), and M ($n = 61$). An ANOVA confirmed that conditions did not differ in prior knowledge as measured by pretest scores, $F(2, 196) = 0.04$, $p = .97$.

Conditions

The study contrasted three conditions: M, TM, and PTM conditions, which differed in how the Coach responded with guidance each time students submitted a solution. The M condition received a motivational message along with “you’re not quite there yet, keep going.” The TM condition received a motivational message plus a task reminder stating a constraint the students’ solution violated. The PTM condition alternated between receiving either a motivational message combined with a task reminder or a problematizing activity. Both TM and PTM conditions received guidance based on an adaptive algorithm that called guidance

[2] Six students (evenly distributed across conditions) were excluded from our sample because they never submitted an invented solution, and received no guidance. Thus, they did not receive the intended treatment of their respective conditions.

Table 1 Design of conditions

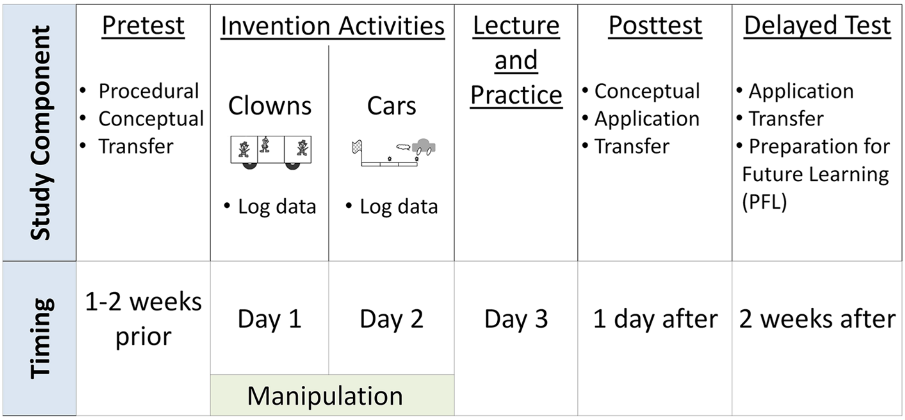
Condition	Problematizing Activities	Task Reminders	Motivational Messages
Problematizing + Task + Motivational Guidance (PTM)	X	X	X
Task + Motivational Guidance (TM)		X	X
Motivational Guidance (M)			X

matched to the submitted solution type. The PTM condition also received additional adaptivity *within* each problematizing activity, which contained a series of prompts and responses from the Coach. All conditions received contrasting cases and had access to tools (calculator, note pad, and rules). Table 1 shows the main components of each condition.

Procedure

Figure 5 shows the study procedure. One to two weeks prior to the start of instruction, students took a paper and pencil pretest. The instruction phase took place on three consecutive days, for 35 min per day. On the first two days, students worked with the Invention Coach software to complete two Invention activities: clown crowdedness and car fastness (i.e. speed). The next day they received a lecture-and-practice session on ratio structures in physics. The following day students took a posttest, and two weeks later they completed a delayed transfer test. Note that the posttest was given after the lecture-and-practice session because the exploratory Invention activities are meant to prepare students to learn from later expository instruction. All activities and instruction were led by one lead researcher, while 3 additional researchers helped to facilitate the study.

During both Invention activities, students worked individually with the version of the Invention Coach that corresponded to their condition. Each Invention task was introduced by a short video explaining the task goals and



	Pretest	Invention Activities		Lecture and Practice	Posttest	Delayed Test
Study Component	• Procedural • Conceptual • Transfer	Clowns • Log data	Cars • Log data		• Conceptual • Application • Transfer	• Application • Transfer • Preparation for Future Learning (PFL)
Timing	1-2 weeks prior	Day 1	Day 2	Day 3	1 day after	2 weeks after
		Manipulation				

Fig. 5 Study procedure highlighting instructional activities and data collected

constraints. The video also mentioned that the task was challenging and would likely require multiple solution attempts.

In all versions of the Invention Coach, students stopped inventing when either they entered the correct set of index numbers or 30 min had passed. At the end of the activity students received a prompt asking them to describe their general solution method in words and with a mathematical equation.

Manipulation of conditions occurred only on days 1 and 2 of instruction, when students worked on Invention activities using their condition's version of the Coach. On day 3 of instruction, all students participated in a whole-class, combined lecture-and-practice session on the importance of ratio structures in science. During a PowerPoint lecture, the lead experimenter revealed the scientifically accurate solution to the crowded clowns Invention task and related crowdedness to the concept of density. Students were presented with the equation for density (density = mass / volume) and were shown a worked example of a density calculation problem. Students were then given an opportunity to practice applying the equation with similar problems on a paper worksheet. Afterwards, the lead experimenter presented an analogous lecture and practice session for the cars task, introducing the concept of speed and its formula and engaging students in practice computing it. The lecture concluded with an explanation of ratio structures. It described the purpose of ratio structures – to compare two quantities with a persistent, inverse relation. The lecture also discussed how ratio-based equations are common in the physical sciences.

Measures

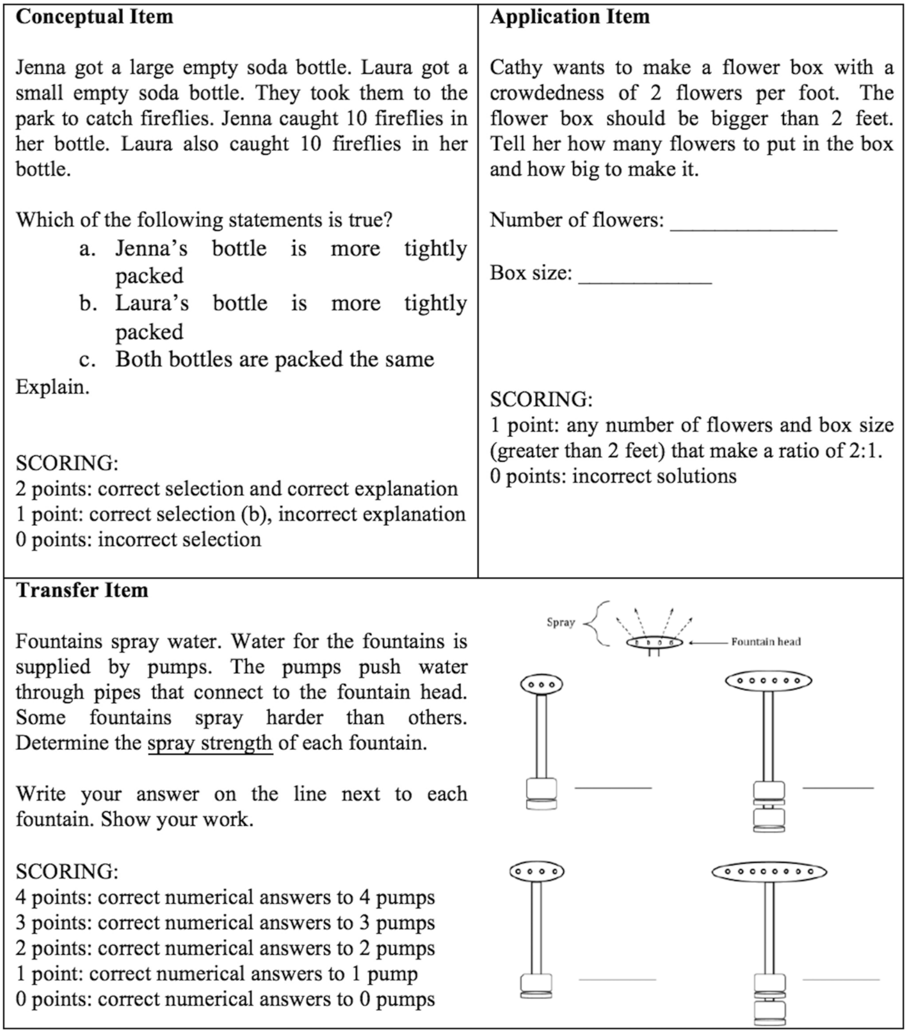
Measures include assessments of learning and transfer, Invention task performance, metrics of exploration and generativity, and measures of the student experience of guidance.

Learning and Transfer

Students took three paper tests which assessed various learning and transfer metrics: a pretest, posttest, and delayed test. Example test items are shown in Fig. 6.

Pretest The pretest contained 6 items: 2 procedural learning, 3 conceptual learning, and 1 transfer item. The procedural learning items assessed whether students had prior knowledge of the equations for density and speed and could apply them accurately (e.g. "Imagine a chocolate heart that is 6 cubic centimeters big. If it has a mass of 48g, what is the density of the chocolate?"). The conceptual learning items targeted students' qualitative understanding of ratio relationships in the context of density and speed, which were explicitly taught in the post-invention lecture. The transfer item, which was identical to one of the posttest transfer items, served as a baseline. An average pretest score was computed, averaging across all items.

Posttest The posttest contained 8 items: 3 conceptual learning, 2 application of learning, and 3 transfer items. The conceptual learning items were isomorphic versions of the pretest items (counterbalanced). Application of learning items



Conceptual Item

Jenna got a large empty soda bottle. Laura got a small empty soda bottle. They took them to the park to catch fireflies. Jenna caught 10 fireflies in her bottle. Laura also caught 10 fireflies in her bottle.

Which of the following statements is true?

a. Jenna's bottle is more tightly packed
b. Laura's bottle is more tightly packed
c. Both bottles are packed the same

Explain.

SCORING:
2 points: correct selection and correct explanation
1 point: correct selection (b), incorrect explanation
0 points: incorrect selection

Application Item

Cathy wants to make a flower box with a crowdedness of 2 flowers per foot. The flower box should be bigger than 2 feet. Tell her how many flowers to put in the box and how big to make it.

Number of flowers: _______________

Box size: ____________

SCORING:
1 point: any number of flowers and box size (greater than 2 feet) that make a ratio of 2:1.
0 points: incorrect solutions

Transfer Item

Fountains spray water. Water for the fountains is supplied by pumps. The pumps push water through pipes that connect to the fountain head. Some fountains spray harder than others. Determine the spray strength of each fountain.

Write your answer on the line next to each fountain. Show your work.

SCORING:
4 points: correct numerical answers to 4 pumps
3 points: correct numerical answers to 3 pumps
2 points: correct numerical answers to 2 pumps
1 point: correct numerical answers to 1 pump
0 points: correct numerical answers to 0 pumps

Fig. 6 Example test items

asked students to reason about density and speed ratios in novel ways (e.g. solving for volume or mass instead of density). Transfer questions assessed whether students could notice and implement ratio structures in novel domains that were not explicitly taught (e.g. pressure, spring constant). Transfer items were adapted from Schwartz et al. (2011). An average item score was computed for each item type on the posttest.

Delayed Test The delayed test was conducted to assess whether any condition differences in application of learning or transfer[3] would remain two weeks after the instruction ended. It contained 2 application and 4 transfer items, which were isomorphic to

[3] Due to limited classroom time, we were unable to include conceptual items on the delayed test.

items on the pretest and posttest. The delayed test also contained 2 additional preparation-for-future-learning (PFL) items (Bransford and Schwartz 1999) that assessed how well students were prepared to learn new ratio-based concepts such as power (p = work/time) and acceleration (a = speed/time). These test items provided some instruction on the test (e.g. a worked example of an acceleration problem, a written passage explaining power) before posing a novel question about the new topic. PFL items constitute a measure of remote transfer, since they differ from the initial learning situation on multiple dimensions (e.g. time, task, content, see Klahr and Chen 2011). An average item score was computed for each item type on the delayed test.

All test items were scored for correctness ranging from incorrect to fully correct (some items had finer-grained scales than others). Item scores were then scaled from 0 to 1. All summative scores represent average item scores (max score = 1).

Each test item was scored by two coders. Reliability for each item was satisfactory, with Cohen's κ ranging from 0.7 to 1.0. All disagreements were discussed and adjudicated. Cronbach's α was 0.33 for the pretest, perhaps due to the small number of pretest items and because students had very little prior knowledge of the subject and may have been guessing the answers. Cronbach's α was 0.58 for the posttest, which is acceptable given the inclusion of diverse item types. Cronbach's α was 0.70 for the delayed test, which is satisfactory. Thus, tests had reasonable internal consistency.

Invention Task Performance

For each Invention task (see Fig. 1 for an example), students received a score of 1 if they invented a completely correct solution (i.e. all 6 index numbers correct). This number was summed across the crowded clowns and cars tasks to create a 0–2 score.

Exploration and Generativity

Log data was collected from all student interactions with the Coach, using the DataShop tools (Koedinger et al. 2010). Each student's logs were used to calculate measures related to their generativity and exploration, derived from learners' submitted solutions. A "solution" is a set of index numbers submitted to the Coach. A "solution type" is a particular class of solution (e.g. unclassifiable, single-feature, two-feature, mathematical two-feature, or ratio-based). Measures of exploration and generativity include the quantity and breadth of solutions generated, rates of solution generation, and solution quality.

Quantity Quantity was measured by the total number of complete solutions submitted. Solutions were considered complete when they contained index numbers for 4 or more cases (out of 6 total).

Breadth The breadth of solutions explored was computed as the total number of viable solution types submitted. There were four viable solution types: one-feature, two-feature, mathematical two-feature, and ratio (excluding the unclassifiable type), yielding a max score of 4.

Rate of Generation We also calculated the rate at which solutions and solution types were generated each minute. The rate of solution generation was computed by dividing the total number of solutions by time spent generating. The rate of solution type generation was computed by dividing the total number of solutions types by time spent generating.

Quality To get a sense for whether students were generating higher quality solution types, we also computed the relative proportions of solutions submitted of each type: unclassifiable, single-feature, two-feature, mathematical two-feature, and ratio-based solutions, which are listed in order from lower to higher quality (see Adaptivity section). Note that ratio-based solutions were in ratio form but were not necessarily the correct solution.

Instructional Experience

To capture the students' experience of the guidance in each condition, we examined students' spontaneous comments on task difficulty, the guidance received in each condition, and time spent on activities within the system.

Difficulty An exploratory post-hoc analysis of reflection prompts revealed interesting data about students' perceptions of task difficulty. In response to a post-invention prompt asking for a description of their general solution method, many students spontaneously mentioned experiencing difficulty during the task, either discussing how challenging the task was (e.g. "it was hard") or how they could not succeed (e.g. "I failed!!!" or "I kept getting the wrong answer"). We coded responses to this question for whether students mentioned difficulty or not. Two coders scored 25% of the data and achieved excellent reliability, κ = .95. The remainder of the data was scored by a single master coder.

Guidance Frequencies were calculated for each of the various forms of guidance students received or accessed. Tool use frequency was computed by summing instances of opening the notepad, rules, and calculator. Frequencies were also generated for motivational messages (M and TM conditions), task reminders[4] (all conditions), and problematizing activities (PTM condition only).

Time We first calculated total time on task, as the time between log in and log out of the system for each Invention task. The mean time spent on the two Invention tasks differed significantly by condition, $F(2, 196) = 4.27$, $p = .02$. Posthoc tests revealed that the PTM condition spent significantly more time in the system compared to the M condition, $M_P = 29.9$, $SD_P = 4.7$; $M_{NP} = 27.6$, $SD_{NP} = 8.0$; $M_M = 26.2$, $SD_M = 8.8$, though the difference is only a few minutes. This is likely because the PTM group was less successful at the Invention tasks. Given this difference in total time, we calculated proportions of times spent on various system components, to assess how students distributed their time in each version of the Invention Coach.

[4] Note that motivational messages and task reminders were given even for incomplete solutions, so the total amount of messages should be greater than the total amount of complete solutions submitted.

Tool use time was calculated by summing time spent in calculator, rules sheet, and notepad. The time spent reading *messages* (motivational messages + task reminders) was inferred from the time a message appeared on the screen until the next user-performed action. We could not distinguish between time spent on motivational messages and task reminders, since they appeared in the same on-screen pop-up. *Activity* time was calculated as the time a problematizing activity started to the time it ended (generally students were locked into an activity once it started and could not access tools or change their index numbers). *Generation* time was defined as the time students were not reading messages, completing activities, or using tools. All time metrics were converted to proportions by dividing by total time on task.

Results

In the following analyses, we use variants of ANCOVA models to explore the effects of condition on continuous, normally distributed outcome measures, and we use ordinal or logistic regression to explore condition effects on ordinal and dichotomous outcomes. Some of the log data constitutes over-dispersed count data, and thus, negative binomial regression (a form of poisson regression) was used to analyze these data. All post-hoc tests use the Bonferonni correction, to correct for Type I error.

In almost all analyses, we controlled for pretest scores and time in the Invention Coach, by including them as covariates in our analysis. We controlled for time because (1) the PTM condition spent slightly more time in the Coach and (2) because students stopped the task once they were successful, time inventing impacts the number of actions in the system. To control for differences across class, a class variable was included as a factor in all analyses (we did not have enough classes to conduct hierarchical linear models).

We explored condition x pretest and condition x time interactions but found none. This suggests that high and low performing students did not respond differently to conditions. Given the lack of interactions, all models contain main effects of variables only.

Four participants were excluded from all log data analyses because the number of solutions they submitted was greater than 3 standard deviations above the mean and these individuals were skewing results within their respective conditions.

Learning and Transfer

Posttest A MANCOVA analyzed condition differences on conceptual, application, and transfer posttest items, using condition and class as factors, while covarying pretest and time. There were significant umbrella effects for all variables, including class, $F(24, 558) = 2.05$, $p = .02$, $\eta_p^2 = .08$, pretest, $F(3, 184) = 10.27$, $p < .001$, $\eta_p^2 = .14$, and time, $F(3, 184) = 8.19$, $p < .001$, $\eta_p^2 = .12$. Of most interest is the significant effect of condition, $F(6, 370) = 2.41$, $p = .03$, $\eta_p^2 = .04$. Follow-up univariate ANCOVAs revealed that conditions only differed on posttest transfer scores, $F(2, 186) = 4.76$, $p = .01$, $\eta_p^2 = .05$. Post-hoc tests on transfer scores showed that the PTM condition outperformed the M condition, $p = .007$, while the TM condition did not differ significantly from other conditions, p's > .27. Thus, by inference, the TM group's transfer score fell somewhere in between the other conditions (see Fig. 7). In sum, the PTM

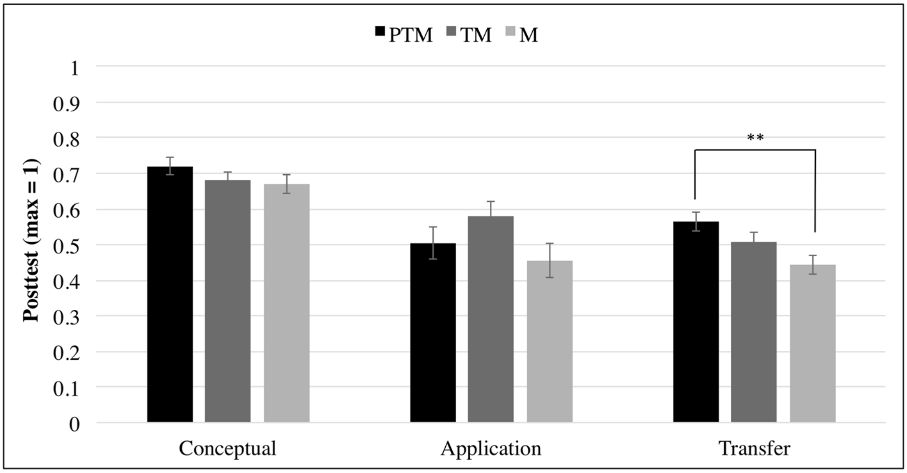


Fig. 7 Adjusted mean scores on posttest items. Error bars = ±1 SE. ** significant difference at $p < .01$

condition had higher posttest transfer scores than the M condition, while the TM group fell somewhere in the middle. However, conditions did not differ in their performance on either conceptual or application posttest items.

To explore the efficacy of the full version of our Invention Coach, we compared the transfer gains made by the PTM condition to those made by students being tutored by humans in a previous study (Chase et al. 2015). We conducted paired *t*-tests on transfer items that were common to pre and posttests, then computed effect size gains. The PTM condition made significant gains on transfer items, $t\ (67) = 5.31$, $p < .001$, an effect size gain of $d = .7$, which is comparable to the transfer gain achieved by students coached by human tutors, $d = .7$.

Delayed Test To explore effects on the delayed test, A MANCOVA was run using application, transfer, and PFL average item scores as outcomes, with condition and class as factors, covarying pretest and time on task. There were significant umbrella effects for class, $F(24,\ 558) = 2.67$, $p < .001$, $\eta_p{}^2 = .10$, pretest, $F(3,\ 184) = 7.47$, $p < .001$, $\eta_p{}^2 = .11$, and time, $F(3,\ 184) = 9.63$, $p < .001$, $\eta_p{}^2 = .14$. However, there was no main effect of condition, $F(6,\ 370) = 0.77$, $p = .59$, $\eta_p{}^2 = .01$. Thus, conditions did not differ in their scores on the delayed posttest items (Fig. 8).

Invention Task Performance

Success on Invention Tasks To test whether conditions differed in their performance on the Invention tasks, we conducted an ordinal regression predicting the total number of successfully solved Invention tasks (0, 1, or 2), using condition as our main predictor, with class and pretest as control variables. Given that students stopped the task once they were successful, resulting in shorter time on task, we did not include time on task as a covariate in this model. Pretest was not a significant predictor of performance, $p = .99$, but several classes had significant effects, p's $< .003$. Dummy variables for TM and M conditions were both significant predictors of performance, $B_{NP} = .85$, *Wald* =

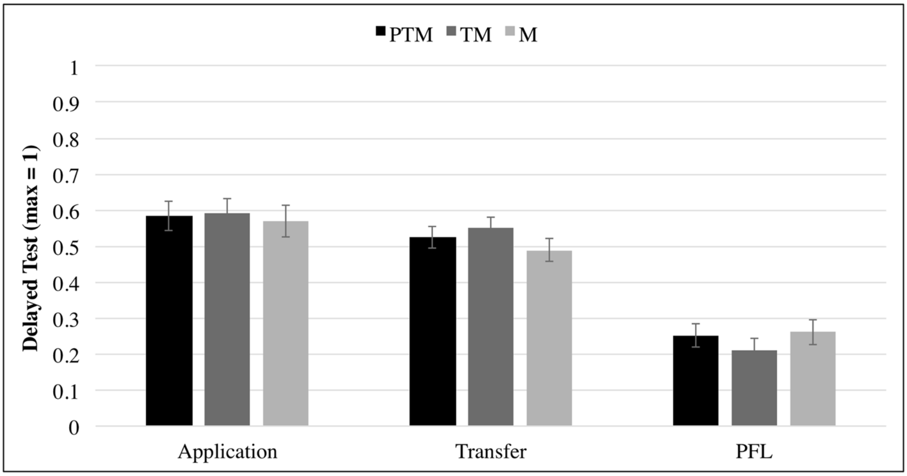


Fig. 8 Adjusted mean scores on delayed test items. Error bars = ±1 SE

4.77, $p = .03$, $B_M = .81$, $Wald = 4.11$, $p = .04$. Thus, being in the M condition led to a 2.34 increase in the odds of performing better, while being in the TM condition led to a 2.25 increase in the odds of performing better, relative to the PTM condition. However, the TM and M conditions did not differ in their performance, $p = .91$. Table 2 shows the percent of students in each condition who invented a correct solution on zero, one, or both Invention tasks. Overall, we found that students in the PTM condition were less likely to invent the correct solution during the problem-solving sessions.

Exploration and Generativity

Quantity To test whether conditions differed in the quantity of solutions they generated, a negative binomial regression was run with the total number of solutions submitted as the outcome variable, using condition as our predictor of interest, with class, pretest, and time included as controls. The model predicted better than chance, χ^2 (12, $N =$ 195) = 48.16, $p < .001$. Significant predictors were class variables, $Wald = 24.36$, $p = .002$ and time, $B = 0.04$, OR = 1.04, $Wald = 35.63$, $p < .001$. However, neither pretest nor dummy variables of condition had significant effects, p's > .16.

Table 2 Percent students inventing the correct solution on 0, 1, or both Invention tasks

	0 tasks	1 task	2 tasks
PTM*	70.6%	26.5%	2.9%
TM	61.4%	17.1%	21.4%
M	60.7%	19.7%	19.7%

*significant condition differences at $p < .05$

Breadth Another measure of exploration is the breadth of solution types generated, or the total number of viable solution types students generated. Since this variable was normally distributed, we explored condition differences with an ANCOVA, while including class as a control factor, and pretest and time as covariates. There were significant effects of time, $F(1, 182) = 13.72$, $p < .001$, $\eta_p^2 = .07$ and class, $F(8, 182) = 3.46$, $p = .001$, $\eta_p^2 = .13$. Pretest did not have a significant effect, $p = .34$. However, condition was a significant factor, $F(2, 182) = 15.15$, $p < .001$, $\eta_p^2 = .14$. Posthoc tests revealed that the PTM condition generated fewer types of solutions than the other two conditions, p's < .02. The TM and M conditions did not differ from one another, $p = 1.0$.

Rate of Generation To test whether conditions differed in the rate at which they generated solutions and solution types, a MANCOVA was run with the rate of solution generation and rate of solution type generation as outcome variables, using condition and class as factors, with time and pretest as covariates. The MANCOVA revealed significant umbrella effects for time, $F(2, 181) = 38.59$, $p < .001$, $\eta_p^2 = .30$, and class, $F(16, 364) = 1.81$, $p = .02$, $\eta_p^2 = .08$. Most importantly, there was a significant condition effect, $F(4, 364) = 4.93$, $p = .01$, $\eta_p^2 = .05$. Individual ANOVAs demonstrated significant condition effects for both rate of solution generation, $p = .005$ and rate of solution type generation, $p < .001$. Posthoc tests revealed that the PTM condition had higher rates of both solution generation and solution type generation compared to both TM and M conditions, p's < .03. Table 3 shows summary statistics of quantity, breadth, and rate of solution generation.

Quality To explore how the quality of solutions differed across conditions, we examined the relative proportions of various types of solutions students generated (Table 4). Since these data were normally distributed, we conducted a MANCOVA with condition and class as factors, pretest and time as covariates, and proportions of each solution type (unclassifiable, 1-feature, 2-feature, mathematical 2-feature) as outcomes. Note that because proportions add to 1, including all solution types in this analysis would violate assumptions of collinearity. Thus we excluded the proportion of ratio-based solutions from our analysis. However, proportions of ratio solutions are reported in Table 4, and it is obvious that these do not differ significantly by condition.

The MANCOVA revealed that time had a significant effect, $F(4, 179) = 36.36$, $p < .001$, $\eta_p^2 = .45$, while class had a near-significant effect, $F(32, 728) = 1.37$, $p = .09$, $\eta_p^2 = .06$, and pretest was not significant, $p = .16$. Importantly, condition was a significant factor, $F(8, 360) = 2.02$, $p = .05$, $\eta_p^2 = .04$. Follow-up univariate

Table 3 Mean number of solutions, solution types, and rates of each, per Invention task (SD)

	# solutions (quantity)	rate of solution generation*	# solution types (breadth)*	rate of solution type generation*
PTM	19.7 (22.9)	2.4 (2.4)	1.7 (0.8)	0.19 (0.08)
TM	22.6 (24.1)	1.3 (2.4)	2.1 (1.0)	0.14 (0.08)
M	19.5 (20.0)	1.1 (2.4)	2.4 (0.8)	0.14 (0.08)

*significant condition differences at $p < .05$

Table 4 Adjusted mean proportion of solution types (with SD)

	unclassifiable*	1-feature	2-feature	math 2-feature	ratio
PTM	.22(.28)	.38(.28)	.15(.24)	.04(.10)	.22(.23)
TM	.29(.28)	.34(.28)	.13(.23)	.02(.10)	.23(.23)
M	.41(.28)	.28(.28)	.08(.23)	.03(.10)	.20(.23)

*significant condition differences at $p < .05$

ANCOVAs revealed that conditions differed on the proportion of unclassifiable solutions only, $F(2, 182) = 6.92$, $p = .001$, $\eta_p{}^2 = .07$. Post-hoc tests revealed that the PTM condition submitted a lower proportion of unclassifiable solutions compared to the M condition, $p = .001$, while other condition comparisons were non-significant, p's > .05.

In sum, conditions did not differ in the quantity of solutions they generated, however the PTM condition explored a narrower breadth of solution types relative to other conditions. The PTM condition also generated solutions at a faster rate, particularly when compared to the M condition. Finally, conditions generated similar proportions of 1-feature, 2-feature, mathematical 2-feature, and ratio-based solutions,[5] but students in the PTM condition generated a lower proportion of "unclassifiable" solutions, relative to the M condition.

Instructional Experience Data

Difficulty For this analysis, we examined the number of students who spontaneously mentioned difficulty during the post-invention task reflection prompts. To test for condition differences, we ran a logistic regression predicting whether or not a student described experiencing difficulty, with condition as our main predictor, controlling for class, pretest, and time on task. The model predicted better than chance, χ^2 (12, N = 199) = 26.24, $p = .01$. Dummy variables for class were not significant predictors, p's > .18. Students who worked longer on the task were more likely to experience difficulty, $B = .04$, $Wald = 3.99$, $p = .05$, such that every additional minute of work on the task led to a 9% increase in the likelihood of experiencing difficulty. More importantly, students in the TM and M conditions were more likely to experience difficulty relative to students in the PTM condition, $B_{NP} = 1.30$, $Wald = 4.27$, $p = .04$, $B_M = 1.96$, $Wald = 9.78$, $p = .002$. In other words, compared to students in the PTM group, students in the TM condition were 3.66 times as likely to report difficulty, and those in the M condition were 7.08 times as likely to report difficulty during the Invention tasks. However, the likelihood of reporting difficulty did not significantly differ across the TM and M conditions, $p = .17$. The raw data demonstrate this pattern as 25% of M students, 17% of TM students, and 6% of PTM students reported experiencing difficulty on at least one of the Invention tasks. Thus, even though

[5] The fact that proportion of ratio-based solutions does not differ by conditions seems at odds with the Invention task performance results, which showed that the PTM group was less likely to generate the correct solution. However, the measure of ratio solutions here counts any form of ratio-based solution, not just correct ones.

Table 5 Mean guidance frequencies per Invention task (with SD)

Condition	Tool Use*	Motivational Messages*	Task Reminders*	Problematizing Activities
PTM	7.1 (4.3)	17.4 (20.0)	17.4 (20.0)	7.9 (4.5)
TM	8.7 (6.3)	27.8 (27.6)	27.8 (27.6)	N/A
M	11.3 (9.6)	22.6 (21.4)	N/A	N/A

*significant condition differences at $p < .05$

students in the PTM condition performed worse on the Invention tasks, they were less likely to spontaneously report experiencing difficulty during the task.

Guidance We then conducted exploratory analyses of log data to get a feel for how frequently students received various forms of guidance as they worked in each version of the Invention Coach (Table 5). To identify these condition differences, we conducted separate negative binomial regressions predicting tool use, motivational messages, task reminders, and guidance cycles, each time using condition as our main predictor, with class, pretest and time included as controls. For the sake of brevity, we report only the results of condition effects for each regression.

We first explored condition differences in tool use. The M condition was associated with more tool use than the PTM condition, $B_M = .60$, $Wald = 9.45$, $p = .002$, such that students in the M condition used the tools 1.82 times more frequently than those in the PTM condition. There were no other significant condition differences, $p's > .08$.

The frequency of motivational messages also differed by conditions. The PTM condition received fewer motivational messages relative to the TM condition, $B_{TM} = .59$, $Wald = 10.67$, $p = .001$ and fewer than the M condition as well, $B_M = .53$, $Wald = 7.02$, $p = .008$. Given that motivational messages co-occur with task reminders in PTM and TM conditions, we can also infer that the PTM group received significantly fewer task reminders than the TM condition.

Overall, we found that conditions differed in frequency of tool use (sum of calculator, notes, and rules sheet access), motivational messages, and task reminders received. Not surprisingly, the M group had more frequent tool use than the PTM condition. The TM condition received more task reminders than the PTM group. The TM and M groups received similar numbers of motivational messages. This suggests that conditions were implemented as intended – the TM group received the most task reminders, the PTM group received fewer reminders but also worked through the Coach's problematizing activities, and the M guidance group was more likely to provide self-help by using the tools.

Time To examine how each condition spent their time in the system, we compared conditions on the proportion of time they spent on each of the main activities within the Invention Coach system: generation, tools, reading messages (combined motivational + task reminder), and problematizing activities.

An initial MANCOVA was run on proportions of generation, tool, and message time, with condition and class as factors, covarying pretest and total time. All effects were significant, but we focus on the condition effect, $F(6, 362) = 53.45$, $p < .001$,

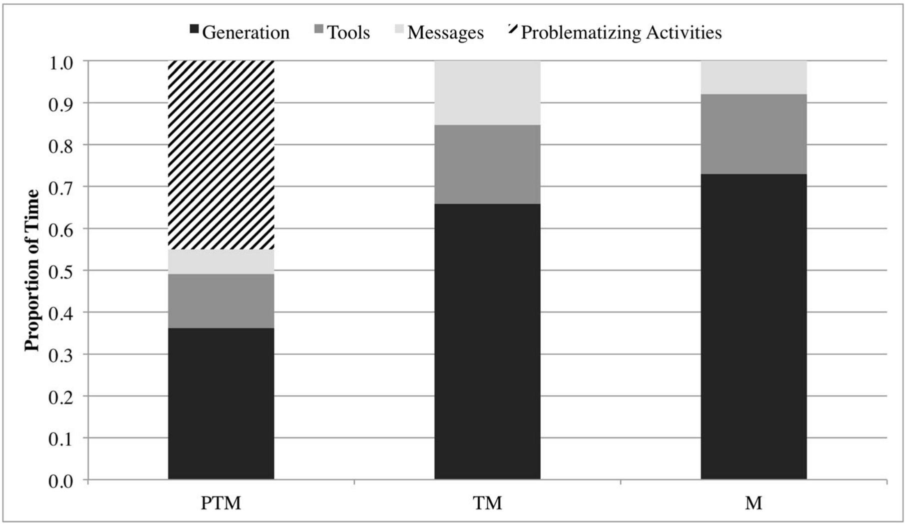


Fig. 9 Proportions of time spent in various components of the Invention Coach

$\eta_p^2 = .47$. Follow-up individual ANCOVAs revealed significant condition effects for generation, $F(2, 182) = 95.79$, $p < .001$, $\eta_p^2 = .51$, tools, $F(2, 182) = 4.85$, $p = .009$, $\eta_p^2 = .05$, and messages, $F(2, 182) = 27.68$, $p < .001$, $\eta_p^2 = .23$. Post-hoc tests revealed that PTM students had significantly lower proportions of generation and tool use time compared to the other two conditions, p's < .03. Also, the TM group spent a greater proportion of time on messages, compared to the PTM and M conditions, $p < .001$.

Overall, we found that the PTM condition spent less time generating solutions, reading messages, or using tools and instead spent this time in problematizing activities. It is interesting to note that from the perspective of how learners spent their time in the system, M and TM conditions look fairly similar, spending roughly the same proportion of time generating solutions, accessing tools, and reading messages. Moreover, the largest proportion of their time was spent generating, followed by tools, then reading messages (according to descriptives). In contrast, the PTM condition looks qualitatively different from the other conditions, by spending the largest proportion of their time in problematizing activities, followed by generation time, tool time, and then messages. Of course, the PTM students spent less time generating and using tools relative to other conditions. Figure 9 demonstrates this pattern of results.

Discussion

We developed the Invention Coach – an intelligent learning environment to support the exploratory phase of Invention, an explore-then-explain (ETE) form of pedagogy. Our full version of the system contains adaptive problematizing guidance that encourages learners confront and grapple with domain complexity and key disciplinary ideas. To test the impact of problematizing guidance, we compared the problematizing Coach (PTM) to two minimally guided versions of the system: a version that provides only motivational support (M) and a version that provides both motivational support and additional clarification of task goals and constraints (TM). Critical outcome measures included learning and transfer of mathematical structures in science and generative and exploratory behaviors. We also explored how task performance and the experience of guidance differed across conditions.

We had hypothesized that the PTM version of the Coach would outperform the minimally guided versions on three posttest measures (conceptual, application, and transfer items). This hypothesis was confirmed with respect to the transfer items only. Across all conditions, students performed similarly on items assessing learning of concepts and their application. However, the PTM version of the Coach facilitated superior transfer, particularly when compared to the most minimally guided version of the Coach (M). The transfer gains achieved by the problematizing version of the Coach were comparable to those accomplished by human tutors in a similar study (Chase et al. 2015). This is a significant achievement, as human tutors are often touted as the "gold standard" in providing individualized, nuanced, and effective guidance to students (Cohen et al. 1982; Lepper et al. 1997). We would also like to point out that we did not find, in any of our analyses, that the effect of guidance was moderated by students' prior knowledge. Thus, we found no evidence that problematizing guidance might be beneficial only for high performers.

Why did problematizing guidance enhance transfer but not learning? This result is not entirely surprising, as some studies have found that Invention tasks enhance transfer but not learning (Schwartz and Martin 2004). It is possible that the different Coach versions were equally effective at facilitating learning, or that most of the learning occurred during the lecture-and-practice session which all conditions received. Alternatively, these results may be due to problematizing's focus on helping learners "attend to and engage with key disciplinary ideas they would otherwise overlook" (Reiser 2004). One of the keys to successful transfer is noticing or perceiving key deep structures (such as ratio) in novel situations (Chase et al. 2019; Day and Goldstone 2012; Gick and Holyoak 1983). However, in non-transfer situations, there are strong cues to help learners notice the relevant concepts (e.g. our learning test items explicitly use the word "density" or "crowdedness"). Thus, perhaps our problematizing guidance was driving students to learn to notice the deep features and structures of ratio, which is particularly important for successful transfer. Another possibility is that the problematizing guidance taught learners effective problem-solving strategies or what others have called inquiry strategies (Holmes et al. 2014), such as critically analysing and contrasting examples, searching for general explanations or models that will work across a set of examples, and reflecting on their solution method, all of which students are being pushed to do in our problematizing activities, and which could contribute to successful performance on our transfer items (which are similar to novel Invention tasks). While other research has found that standard unguided Invention activities do not lead learners to transfer problem-solving strategies (Schwartz et al. 2011), it is possible that our problematizing guidance had this effect.

Unfortunately, transfer effects did not hold up on a 2-week delay. However, we believe this result is inconclusive since the delayed test was compromised. After the study was completed, teachers revealed that students had received instruction on ratio and speed problems in their math classes between the post and delayed tests. Moreover, when coding students' answers to delayed test questions, some students used common solution methods and terms that did not appear on the posttest and were not part of our instruction (e.g. comparing ratios using a common unit). Our interpretation of the delayed test results is that condition differences may have been washed out by the incidental instruction students received during the delay. However, it is also possible that the effects of the problematizing version of the Invention Coach simply did not last.

Despite their superior transfer, the PTM condition performed worse on the Invention tasks (were less likely to invent the correct solutions), relative to more minimally guided conditions. On reflection, this finding makes sense for a number of reasons. First, in ETE paradigms, there is no strong expectation that students will invent the correct solution, only that attempting to invent a solution will prepare students to learn from later expositions. Some ETE studies have demonstrated that high performance on the exploratory task is not necessary for transfer to occur (Schwartz and Martin 2004). Second, problematizing can “make things more difficult in the short run but be productive for learning” (Reiser 2004). Thus, instead of spending their time working out correct solutions, these students were working on problematizing activities, grappling with the deep ideas that underlie the domain. Moreover, this finding is consonant with related research on productive failure (Kapur 2008) and desirable difficulties (Bjork 1994), which suggest that difficulties during learning experiences may actually facilitate later learning, or in this case, transfer. Thus, it is entirely plausible that while students in the PTM condition performed poorly on the Invention task relative to other conditions, they displayed superior transfer.

While students in the PTM condition were less successful on the Invention tasks, they were also less likely to spontaneously report experiencing difficulty during the tasks. On the one hand, we found this surprising, since one potential danger of making an aspect of the situation “problematic” for students is that it could also make the subjective experience of the task more frustrating. On the other hand, the problematizing activities also provide an added degree of structure (which is difficult to separate from problematizing), in that the activities contain several prompts and decisions that are designed to focus learners on the relevant content. If students are able to succeed at these simpler subtasks, then they may feel as if they are making progress on the task or mastering the content, which could reduce the experience of difficulty within the Invention task. Nonetheless, the problematizing guidance succeeded in mitigating students' perceptions of task difficulty on the inherently challenging Invention tasks.

We also investigated how these forms of guidance would shape learners’ exploratory and generative activity in the Invention tasks. One reason we chose problematizing guidance for the “full version” of our Coach is because we predicted it would not reduce productive exploration, as other forms of highly structured, direct guidance might (Schwartz et al. 2011). We found that even though students in the PTM condition spent less *time* generating solutions, they generated the same *quantity of solutions* compared to students in more minimally guided conditions (about 20 per task!). This is because PTM students generated solutions at a faster rate than students who received other forms of guidance. Moreover, relative to other conditions, PTM students generated the same proportion of high quality solutions (e.g. ratio-based and mathematical 2-feature solutions) but fewer random guesses (e.g. unclassifiable solutions). We also found that PTM students generated fewer solutions types, meaning they explored less broadly. Thus, one interpretation of the problematizing group's behavior is that they were exploring the solution space less broadly but in a more targeted way, possibly avoiding trial-and-error approaches, which partially confirms our hypothesis.

Despite this fairly productive exploration, students in the PTM condition were still less likely to produce the correct ratio solution on the Invention tasks, which is curious. One possibility is that for some students, the problematizing guidance pushes them towards productive exploration, while for others, it may confuse or distract them. While we did not find interactions between condition and prior knowledge in any of our analyses, it is possible that for students who are able to problematize and explore deeply on their own, the problematizing guidance slowed down or hindered their problem-solving process in some way.

An interesting pattern in the results is that that the two minimal guidance conditions either performed similarly or the TM condition performed somewhere in the middle, in between the PTM and M conditions, on most measures. This suggests that task reminders had some positive, though not always significant, effects on exploration, experience of difficulty, and transfer. Thus, adding clarification of the goals and constraints of the Invention task – pointing out the constraint a solution violated – only slightly helped learners. On the other hand, learners in the TM and M conditions apportioned their time in the system in similar ways (e.g. lots of time generating and using self-help tools), which was quite different from the PTM condition. In this regard, the learning experience of the TM and M conditions was fairly similar, and this may be why differences between these two conditions on most measures were small to none.

Implications

This study demonstrated that the problematizing guidance provided by the Invention Coach engaged students in more productive exploration that was both faster and more targeted towards deep domain features and structures. Thus, it would seem that adaptive problematizing styles of guidance can effectively support Invention without quelling the exploratory nature of the task, and in fact, the guidance seems to enhance the quality and rate of solution generation. Moreover, this work demonstrated that our brand of problematizing guidance was effective in enhancing transfer in the short term while reducing the experience of difficulty during the task. Thus, problematizing guidance is a promising method of support for exploratory problem-solving activities, particularly those that follow an ETE pattern. This work also adds to the small body of work supporting the efficacy of problematizing guidance (Efstathiou et al. 2018; Hicks and Doolittle 2008; Molenaar et al. 2011) and the even smaller body of work exploring adaptive problematizing guidance (Efstathiou et al. 2018).

Moreover, while the construct of *problematizing guidance* is broad and open-textured, our work provides a specific example of one brand of adaptive problematizing guidance in the form of interactive activities. While the tell-me-how activity is similar to other forms of structured explanation tools that force learners to explain in the semantics of the discipline (e.g. ExplanationConstructor), the contrast-focused ranking and feature contrast activities which are sometimes used to support Invention (Holmes et al. 2014) present a form of problematizing guidance that was not mentioned in Reiser's (2004) framework and is unique amongst technology-based scaffolds designed for the express goal of problematizing (Efstathiou et al. 2018; Hicks and Doolittle 2008; Molenaar et al. 2011).

In addition, we have generated three broad design guidelines for developing problematizing guidance for ETE activities: (1) refrain from explicit telling forms of guidance (2) use instructional strategies, such as contrast and explain, that reveal knowledge gaps and focus learners on deep ifeatures (3) draw on instructional strategies used by real teachers. These guidelines are provisional, and it remains to be seen how well they will generalize to other systems. At minimum, we believe they can be applied to design guidance for the exploratory phase of other forms of ETE instruction, where the goal is to facilitate productive exploration of a domain.

Finally, research findings on the efficacy of guidance for the explore phase of ETE instruction is mixed, with some scholars advocating for guidance that promotes task success (Wood et al. 1976), limiting guidance to make space for task failure (Kapur 2011), or guiding learners to engage in inquiry around those failures (Holmes Holmes et al., 2014). Our work suggests that guidance which promotes deep exploration of the underlying concepts by problematizing learners' understanding can enhance learner's ability to transfer.

Limitations and Future Directions

One limitation of this work is that we have not isolated exactly what makes the problematizing version of the Invention Coach effective in comparison to the motivational guidance condition. We have argued that problematizing guidance is particularly well-suited for exploratory problem-solving where the goal is to encourage learners to encounter and struggle with deep domain ideas. However, it is possible that our results are due to differences in the amount of guidance or interactivity between conditions. In this first study, we provided an existence proof that our designed guidance could be efficacious relative to minimal forms of guidance. Future studies could identify the precise ingredients of effective problematizing guidance by isolating the effects of adaptivity, amount of guidance, type of guidance, the mixture of problematizing and structuring, and so on.

A second limitation is that our delayed test measure was likely compromised to the point that we cannot draw valid conclusions from it. It will be important to conduct future studies to explore the long-term effects of problematizing guidance on learning and transfer, in addition to replicating the effects described here. Likewise, our exploratory analysis revealed that problematizing guidance may have reduced learners' perceptions of task difficulty. In future work, it will be important to assess the influence of motivational and affective components on students' exploratory problem-solving (but see Lamnina & Chase, 2018).

A third limitation of this work is that the current Coach is only equipped to support invention of ratio-based equations. The Invention process itself has had good results with a fairly wide range of equation types (Holmes et al. 2014; Kapur 2008; Schwartz et al. 2011). In future work, we aim to adapt the Coach to support invention of a broader variety of equation types (additive, multiplicative, exponential, etc.). There is every reason to believe that the same problematizing activities (Ranking, Feature Contrast, Tell-Me-How) will prove useful, as variants of them have been used effectively with paper-and-pencil versions of Invention (Roll et al. 2012) and in one other computerized system (Holmes et al. 2014). Generalizing our current implementation would require substantial redesign and re-implementation. Our development process involved observing one-on-one sessions of experienced teachers guiding students, pilot testing Wizard-of-Oz prototypes (Marks et al. 2016), identifying categories of student-generated solutions, then generating and testing a solution classification algorithm and branching structure for the interactive dialogue. We could imagine undertaking a similar development process to expand the Coach's capabilities beyond simple ratio equations. A different approach would be to develop a more elaborate set of solution categories, which might enable us to build a more generalized algorithm for solution classification.

Conclusion

We have created the first technology that was designed to adaptively problematize learners' understanding while they engage in Invention activities: The Invention Coach. This work suggests that problematizing guidance of explore-then-explain activities can support short-term transfer, reduce subjective task difficulty, and provoke targeted and efficient exploration of the domain space. This research also demonstrates the efficacy of the full version of the Coach, while adding to a small body of evidence suggesting that technology-based guidance designed explicitly to problematize can be effective. This work also provides a provisional set of design guidelines for developing guidance that problematizes, which other researchers and developers could apply. Finally, problematizing guidance shows potential as a viable way to guide other forms of exploratory problem-solving tasks, where a major goal is to help learners explore productively, so they are better prepared to transfer from later expository instruction.

Acknowledgments We thank Jenna Marks for helping to design the Invention Coach and Octav Popescu and Martin van Velsen for building it. This work was funded by the National Science Foundation (award # 1361062).

References

Aleven, V., McLaren, B. M., Sewall, J., van Velsen, M., Popescu, O., Demi, S., … Koedinger, K. R. (2016a). Example-tracing tutors: Intelligent tutor development for non-programmers. *International Journal of Artificial Intelligence in Education, 26*(1), 224–269. https://doi.org/10.1007/s40593-015-0088-2.

Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2016b). Instruction based on adaptive learning technologies. Handbook of research on learning and instruction. Routledge.

Aleven, V., Connolly, H., Popescu, O., Marks, J., Lamnina, M., & Chase, C. (2017, June). An adaptive coach for invention activities. In *International conference on artificial intelligence in education* (pp. 3–14). Cham: Springer.

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology, 103*(1), 1–18.

Biswas, G., Leelawong, K., Schwartz, D., Vye, N., & The Teachable Agents Group at Vanderbilt. (2005). Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence, 19*(3–4), 363–392.

Bjork, R. A. (1994). Memory and metamemory considerations in the. Metacognition: Knowing about knowing, 185.

Bransford, J. D., & Schwartz, D. L. (1999). Chapter 3: Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education, 24*(1), 61–100.

Bransford, J. D., Franks, J. J., Vye, N. J., & Sherwood, R. D. (1989). New approaches to instruction: Because wisdom can't be told. Similarity and analogical reasoning, *470*, 497.

Chase, C. C., Chin, D. B., Oppezzo, M. A., & Schwartz, D. L. (2009). Teachable agents and the protégé effect: Increasing the effort towards learning. *Journal of Science Education and Technology, 18*(4), 334–352.

Chase, C. C., Marks, J., Bernett, D., Bradley, M., & Aleven, V. (2015, June). Towards the development of the invention coach: A naturalistic study of teacher guidance for an exploratory learning task. In *International Conference on Artificial Intelligence in Education* (pp. 558–561). Springer, Cham.

Chase, C. C., Malkiewich, L. M., & Kumar, A. (2019). Learning to notice science concepts in engineering activities and transfer situations. *Science Education, 103*(2), 440–471.

Chen, W., Looi, C.-K., & Wen, Y. (2011). A scaffolded software tool for L2 vocabulary learning: GroupScribbles with graphic organizers. In H. Spada, G. Stahl, N. Miyake & N. Law (Eds.), Proceedings of Computer Supported Collaborative Learning (CSCL) 2011 (Part 1, pp. 414–421). Hong Kong: International Society of the Learning Sciences.

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

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

Cohen, P. A., Kulik, J. A., & Kulik, C. L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19*(2), 237–248.

Collins, A., Warnock, E. H., Aiello, N., & Miller, M. L. (1975). Reasoning from incomplete knowledge. In *Representation and understanding* (pp. 383–415). Morgan Kaufmann.

Day, S. B., & Goldstone, R. L. (2012). The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist, 47*(3), 153–176.

Efstathiou, C., Hovardas, T., Xenofontos, N. A., Zacharia, Z. C., Anjewierden, A., & van Riesen, S. A. (2018). Providing guidance in virtual lab experimentation: The case of an experiment design tool. *Educational Technology Research and Development, 66*(3), 767–791.

Feng, C. Y., & Chen, M. P. (2014). The effects of goal specificity and scaffolding on programming performance and self-regulation in game design. *British Journal of Educational Technology, 45*(2), 285–302.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*(1), 1–38.

Hicks, D., & Doolittle, P. E. (2008). Fostering analysis in historical inquiry through multimedia embedded scaffolding. *Theory & Research in Social Education, 36*(3), 206–232.

Holmes, N. G., Day, J., Park, A. H., Bonn, D. A., & Roll, I. (2014). Making the failure more productive: Scaffolding the invention process to improve inquiry behaviors and outcomes in invention activities. *Instructional Science, 42*(4), 523–538.

Kapur, M. (2008). Productive failure. *Cognition and Instruction, 26*(3), 379–424.

Kapur, M. (2010). Productive failure in mathematical problem solving. *Instructional Science, 38*(6), 523–550.

Kapur, M. (2011). A further study of productive failure in mathematical problem solving: Unpacking the design components. *Instructional Science, 39*(4), 561–579.

Kapur, M. (2014). Productive failure in learning math. *Cognitive Science, 38*(5), 1008–1022.

Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences, 21*(1), 45–83.

Klahr, D., & Chen, Z. (2011). Finding one's place in transfer space. *Child Development Perspectives, 5*(3), 196–204.

Koedinger, K. R., Baker, R. S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. Handbook of educational data mining, 43, 43–56.

Lamnina, M. & Chase C.C. (2018, April) How different types of uncertainty affect learning, transfer, curiosity, and affect. Poster presented at the annual meeting of the American Educational Research Association, New York, N.Y.

Lepper, M. R., Drake, M. F., & O'Donnell-Johnson, T. (1997). Scaffolding techniques of expert human tutors. In K. Hogan & M. Pressley (Eds.), *Advances in learning & teaching. Scaffolding student learning: Instructional approaches and issues* (pp. 108–144). Cambridge: Brookline Books.

Linn, M. C., Bell, P., & Davis, E. A. (2004). *Internet environments for science education*. Mahwah: Lawrence Erlbaum Associates, Inc.

Loibl, K., & Rummel, N. (2014). The impact of guidance during problem-solving prior to instruction on students' inventions and learning outcomes. *Instructional Science, 42*, 305–326.

Loibl, K., Roll, I., & Rummel, N. (2017). Towards a theory of when and how problem solving followed by instruction supports learning. *Educational Psychology Review, 29*(4), 693–715.

Lynch, C., Ashley, K. D., Pinkwart, N., & Aleven, V. (2009). Concepts, structures, and goals: Redefining ill-definedness. *International Journal of Artificial Intelligence in Education, 19*(3), 253–266.

Manu Kapur, (2012) Productive failure in learning the concept of variance. Instructional Science 40 (4):651–672

Marks, J., Bernett, D., & Chase, C. C. (2016). The invention coach: Integrating data and theory in the Design of an Exploratory Learning Environment. *International Journal of Designs for Learning, 7*(2).

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist, 59*(1), 14–19.

Molenaar, I., van Boxtel, C. A., & Sleegers, P. J. (2011). Metacognitive scaffolding in an innovative learning arrangement. *Instructional Science, 39*(6), 785–803.

Piaget, J. (1977). *The development of thought: Equilibration of cognitive structures*. New York: Viking.

Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *Journal of the Learning Sciences, 13*(3), 273–304.

Roll, I., Holmes, N. G., Day, J., & Bonn, D. (2012). Evaluating metacognitive scaffolding in guided invention activities. *Instructional Science, 40*(4), 691–710.

Sandoval, W. A. (2003). Conceptual and epistemic aspects of students' scientific explanations. *The Journal of the Learning Sciences, 12*(1), 5–51.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction, 16*(4), 475–5223.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction, 22*(2), 129–184.

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology, 103*(4), 759–775.

Shemwell, J. T., Chase, C. C., & Schwartz, D. L. (2015). Seeking the general explanation: A test of inductive activities for learning and transfer. *Journal of Research in Science Teaching, 52*(1), 58–83.

Shute, V. J., & Zapata-Rivera, D. (2012). Adaptive educational systems. *Adaptive technologies for training and education, 7*(27), 1–35.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education, 16*(3), 227–265.

Wiedmann, M., Leach, R. C., Rummel, N., & Wiley, J. (2012). Does group composition affect learning by invention? *Instructional Science, 40*(4), 711–730.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Catherine C. Chase[1] · **Helena Connolly**[1] · **Marianna Lamnina**[1] · **Vincent Aleven**[2]

[1] Teachers College, Columbia University, 525 W. 120th St., New York, NY 10027, USA

[2] Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, USA